Journal of Modern Applied Statistical Methods

Volume 14 | Issue 2 Article 20

11-1-2015

Caution for Software Use of New Statistical Methods (R)

Akiva J. Lorenz
Dallas Independent School District, Dallas, TX, akiva@wayne.edu

Barry S. Markman
Wayne State University

Shlomo Sawilowsky
Wayne State University, professorshlomo@gmail.com

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation
Lorenz, Akiva J.; Markman, Barry S.; and Sawilowsky, Shlomo (2015) "Caution for Software Use of New Statistical Methods (R)," *Journal of Modern Applied Statistical Methods*: Vol. 14 : Iss. 2 , Article 20.
DOI: 10.22237/jmasm/1446351540

Journal of Modern Applied Statistical Methods
November 2015, Vol. 14, No. 2, 275-281.
Copyright © 2015 JMASM, Inc.
ISSN 1538 − 9472

Statistical Software Applications & Review
Caution for Software Use of New Statistical Methods (R)

Akiva J. Lorenz
Dallas Independent School Dist.
Dallas, TX

Barry S. Markman
Wayne State University
Detroit, MI

Shlomo S. Sawilowsky
Wayne State University
Detroit, MI

Open source programming languages such as R allow statisticians to develop and rapidly disseminate advanced procedures, but sometimes at the expense of a proper vetting process. A new example is the least trimmed squares regression available in R's lqs() in the MASS library. It produces pretty regression lines, particularly in the presence of outliers. However, this procedure lacks a defined standard error, and thus it should be avoided.

Keywords: R, lqs(), least trimmed squares regression

Introduction

As new methods appear software vendors race to disseminate them, providing a competitive edge in increasing sales. In the past half century there were numerous examples where this led to the inclusion of procedures that were inappropriate or destructive. For example, consider the mainframe version of SPSS's general linear model command in the 1980's. Option 9, a contrast coding least squares regression approach due to Overall and Spiegel (1969), was subsequently shown to test no known statistical hypothesis (see, e.g., Blair & Higgins, 1978a; Blair, 1978; Blair & Higgins, 1978b). Another example is the R aov() function "when conducting analysis of covariance" which "does not work correctly" (Schumacker, 2015, p. 288).

Dr. Lorenz is an Evaluation Analyst with the Dallas Independent School District. Email him at akiva@wayne.edu. Dr. Markman is a Professor in the College of Education. Email him at barry.markman@wayne.edu. Dr. Sawilowsky is a Professor in the College of Education. Email him at shlomo@wayne.edu.

One of many modern approaches to regression is the least squared trimmed means, where the sum of squared residuals are replaced with the "sum of the *q* smallest squared residuals, where *q* is roughly *n*/2" (Verzani, 2004, p. 100). Hence, this is essentially an M (maximum likelihood) estimator. It is invoked in R via the lqs() function located in the MASS package.

Rousseuw and Leroy (1987) indicated least trimmed means regression is resistant to outliers (see also Verzani, 2004, p. 100). Ripley (2004) noted that least trimmed *squares* is based on minimizing "the sum of squares for the smallest *q* of the residuals," where q takes on various values (e.g., S+ and R sets *q* to 90% as the default). The result is a regression model that "maximizes accuracy to the *q*% of data. The quantile squared residual... [with] floor((*n* + *p* + 1)/2)" (Ripley, n.d.), where *n* are data points and *p* are the regressors. lqs() is exact with one regressor. (For further details, see Fox, 2002. Note that this method is ill equipped to recover if there are no outliers, when ordinary regression should have been used. Once data are trimmed, they are removed from further calculations whether they should have been eliminated or not.)

Unfortunately, the lqs() function is not associated with a defined standard error. (This is a common problem with maximum likelihood applications. For example, see Holford, (2002, p. 45) regarding a 2×2 table with zero frequencies in a cell). Hence, the purpose of this study is demonstrate this concern with respect to lqs().

Methodology

The number of repetitions per experiment was 100,000, conducted on an Intel Sandy Bridge i7-2600K 3.4GHz CPU-based computer, with ultra-high speed Corsair Vengeance Low Profile 4x4GB RAM, Crucial M4 256GB solid state hard drive, and the Windows 7 Ultimate 64 bit operating system. This equipment was necessary due to the well known lack of speed of the R platform, and even so the results compiled in each table took more than 45 minutes to complete. Data were produced using R rnorm(). To determine the veracity of the coding, the normal theory ordinary least squares method was used for comparison using R's lm().

Standard error of beta and the lqs() method.

The t test is defined as beta divided by the standard error of beta (Brase & Brase, 2013, p. 536; Mann, 1995, p. 667), which is then associated with $df = N - 2$ for the t (or Z for large samples) distribution. It is generally not optimal to use the

normal theory formula for the standard error (i.e., the standard deviation divided by the sample size), because it is not robust to non-normality distributed. (There are potential alternatives, such as the Winsorized sample standard deviation, or a jackknife or bootstrap approximation. See, e.g., Sawilowsky & Fahoome, 2003, p. 22, 376 - 382. However, there are limitations to those alternatives.)

Wilcox (1996) provided alternatives in computing the standard error for other hypothesis tests (e.g., the sample median), but that was only after a test was presented using the robust estimator in the numerator combined with the normal curve theory standard error in the denominator (see, e.g., p. 120). The same approach could be used here, with the *p*-value associated with beta obtained from lqs() determined via the normal curve theory standard error (i.e., which is produced by the lm() routine).

Results

Using the standard error under lm(y ~ x) (i.e., beta associated with the ordinary least squares regression) as the denominator for the test of beta obtained from lqs() was found to be unsatisfactory, with inflated Type I errors from between 7.3 and 104 times nominal alpha, as noted in Table 1 below.

Table 1. Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	**lqs()**
0.050	0.04972	**0.36455**
0.010	0.01041	**0.21966**
0.001	0.00102	**0.10248**

Note: Values in bold exceed Bradley's (1978) liberal definition of robustness.

An attempt was made to improve the standard error used in lqs() by replacing the original *y* values with the fitted values of *y* obtained from lqs(). The standard error of the estimate (SE_E, or residual standard deviation) was based on

$$SE_E = \sqrt{\sum_i^n \frac{\left(y - y'\right)^2}{n-2}} \quad (1)$$

where *y*' was obtained as fitted values from lqs() instead of the fitted values from lm(). The standard error of beta (SE_b) is determined by

$$SE_b = \frac{SE_E}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \tag{2}$$

Assembling the t test on beta as a ratio of beta divided by (2),

$$t = \frac{b}{SE_b} \tag{3}$$

the obtained t is significant if

$$\left|t_{obt}\right| \geq t_{\frac{\alpha}{2},n-2}$$

Although as noted in Table 2 there was improvement in the Type I error rates, the inflation was nevertheless from between 5.8 and 39.4 times nominal alpha, which is not acceptable. (Note the values for lm() differed slightly from those in Table 1 above due to the change in the seed number).

Table 2. Type I error rates for $n_1 = n_2 = 30$; $r = 0.0$; 100,000 repetitions

α	Test	
	lm()	**lqs()**
0.050	0.05029	**0.29371**
0.010	0.01061	**0.14499**
0.001	0.00109	**0.04151**

Note: Values in bold exceed Bradley's (1978) liberal definition of robustness.

Regarding the least median squares (lms) option (i.e., "method = lms" option in lqs(), which can be used to invoke a variety of robust methods), subsequent to a Monte Carlo simulation Paranagama (2010) concluded, "In practice, the use of LMS is limited by the absence of formulas for standard errors" (p. 35). This difficulty applies to the default method (least trimmed squares), and hence, lqs() must be abandoned if the purpose of conducting the linear model is to

compute a t test on beta until an adequate standard error for the least squares regression algorithm can be found.

Conclusion

An appropriate standard error has not been derived for the lqs() method. Because the t test on β requires the standard error, various options were considered: (1) the *p*-value associated with β obtained from lqs() was determined via the normal curve theory standard error via the lm() procedure, which failed because it produced Type I errors as large as 104 times nominal α, and (2) the standard error was obtained by replacing the original *y* values with the fitted values of *y* obtained from lqs(), which was an improvement, but also failed because it produced Type I errors as large as 39.4 times nominal α.

The lqs() procedure produces pretty regression equations, and visually fits data in situations with outliers better than the normal theory lm(). However, the absence of a defined standard error precludes its usage in practice. Moreover, the method is not even being close to maintaining nominal alpha. The matter will become increasingly serious as applied researchers continue to be attracted to its highly publicized robustness regression lines, ease of availability in R, and implement it in applied work. For example, lqs() was used by Fan, Lu, Madnick, and Cheung (2001) in a study on data integration in information systems, Abo-Khalil and Abo-Zied (2012) in a study of sensorless control of wind turbines, and Gidnaa and Domínguez-Rodrigo (2013) in a study of human femoral length from fragmented specimens.

In conclusion, new methods should be avoided until such time that they are fully vetted. If this caution was true in the past with expensive, major commercial software such as SPSS, then how much more so caution should be invoked when using free, open source software such as R.

References

Abo-Khalil, A. G., & Abo-Zied, H. (2012). Sensorless control for DFIG wind turbines based on support vector regression, *IECON 2012 - 38th Annual Conference on IEEE Industrial Electronics Society*, 3475 - 3480. doi:10.1109/IECON.2012.6389341

Blair, R. C. (1978). I've been testing some statistical hypotheses: Can you guess what they are?, *The Journal of Educational Research*, *72*(2), 116-118.

Blair, R. C., & Higgins, J. J. (1978a). Comments on "Contrast coding in least squares regression analysis," *American Educational Research Journal*, *15*(1), 149-151.

Blair, R. C., & Higgins, J. J. (1978b). Tests of hypotheses for unbalanced factorial designs under various regression/coding method combinations. *Educational and Psychological Measurement*, *38*, 621-631.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144-152.

Brase, C. H., & Brase, C. P. (2013). *Understanding basic statistics*. 6th ed. Boston, MA: Brooks/Cole, Cengage Learning.

Fan, W, Lu, H., Madnick, S. E. & Cheung, D. W.-L. (2001). Discovering and reconciling value conflicts for numerical data integration. *Information Systems*, *26*(8), 635–656. doi:10.1016/S0306-4379(01)00043-6

Fox, J. (2002). Robust Regression. *Appendix to An R and S-Plus Companion to Applied Regression. R-Project.org*. Retrieved from http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-robust-regression.pdf.

Gidnaa, A. O., & Domínguez-Rodrigo, M. (2013). A method for reconstructing human femoral length from fragmented shaft specimens. *HOMO - Journal of Comparative Human Biology*, *64*(1), 29–41.

Holford, T. R. (2002). *Multivariate methods in epidemiology*. Oxford: Oxford University Press.

Mann, P. S. (1995). *Introductory statistics*. 2nd ed. NY: Wiley.

Overall, J. E., & Spiegel, D. K. (1969). Concerning least squares analysis of experimental data. *Psychological Bulletin*, *72*(5), 311-322. doi:10.1037/h0028109

Paranagamap, T. D. (2010). *A simulation study of the robustness of the least median of squares estimator of slope in a regression through the origin model* (Unpublished Master Thesis). Manhattan, KS: Kansas State University. Retrieved from http://krex.k-state.edu/dspace/bitstream/handle/2097/7045/ThilankaParanagama2010.pdf

Ripley, B. D. (n.d.). *Resistant Regression*. Retrieved from http://astrostatistics.psu.edu/su07/R/html/MASS/html/lqs.html

Ripley, B. D. (2004). *Robust statistics*. University of Oxford. Retrieved from http://www.stats.ox.ac.uk/pub/StatMeth/Robust.pdf

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. NY: Wiley.

Sawilowsky, S. S., & Fahoome, G. (2003). *Statistics through Monte Carlo simulation with Fortran*. Rochester Hills, MI: JMASM.

Schumacker, R. E. (2015). *Learning statistics Using R*. Los Angeles: Sage.

Verzani, J. (2004). *Using R for introductory statistics*. Boca Raton: Chapman & Hall/ CRC.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.