An Application of a New Multivariate Resampling Method to Multiple Regression

Haiyan Bai
University of Central Florida

Wei Pan
University of Cincinnati

This study presents a new multivariate resampling method to improve the performance of multiple regression with small samples. The kernel resamping technique (KRT) is utilized in the multivariate resampling procedure to draw random resamples with random noises, which facilitates obtaining more accurate parameter estimates and their standard errors in multiple regression. The findings from an empirical example suggest that the statistical performance of multiple regression is improved through the KRT technique.

Multiple regression is one of the popular statistical methods; however, when it is applied to small samples, it encounters problems related to the accuracy of statistical estimation (Allison, 1999). Resampling method has been implemented to solve the small sample problems (Bai & Pan, 2008; Davison & Hinkley, 1997; Efron & Tibshirani, 1998). The bootstrap as the most popular resampling method has been applied in regression analysis since Efron's pioneer work in 1979. The bootstrap is an effective tool to solve the small sample problems, and comprehensive applications of the bootstrap to regression models have also been developed by researchers (e.g., Bickel & Freedman, 1981, 1983; Freedman, 1981; Peters & Freedman, 1984; Shao, 1988; Weber, 1984; Wu, 1986); however, the bootstrap standard errors tend to be biased downward in regression analysis when applying to a small sample (Peters & Freedman, 1984). Therefore, managing small sample problems in regression still remains a pertinent issue. A study on improving the statistical performance of multiple regression with small samples would significantly contribute to the literature in both methodological research of small sample issues and applied research using multiple regression with small samples.

The purpose of this present study is to introduce a new multivariate resampling method, the kernel resampling technique (KRT), to tackle the problems in multiple regression with small samples. Specifically, the present study (a) introduces the procedure of KRT, (b) examines the performance of KRT in multiple regression through an empirical example, and (c) compares the performance of KRT in multiple regression with that of the bootstrap in terms of estimation bias and standard errors.

Kernel Resampling Technique

Kernel Resampling Technique (KRT) is a new resampling method which uses kernel smoothing technique to capture the shape of the empirical sample distributions and sampling from the neighborhoods of the original sample. The kernel technique uses kernel probability density estimation to map the original linear or non-linear observations into a higher-dimensional space, where the linear classifier is subsequently used to solve problems (Aizerman, Braverman, & Rozonoer, 1964). The kernel probability density estimation has gained popularity, especially for dealing with nonparametric issues (Towers, 2002).

The kernel technique has been used in the bootstrap for smoothing the bootstrap distribution (Efron & Tibshirini, 1998; Silverman & Young, 1987). However, it is worth noting that there are two key points when using the kernel technique in the smooth bootstrap: (1) kernels are used *after* the bootstrap resampling to smooth the bootstrap distribution, and (2) the bandwidths of the kernels are not specifically defined for different data in the smooth bootstrap. With regards to the second point, finding an optimal bandwidth for the bootstrap smoothing procedure is a statistically and technically challenging task for many researchers and statisticians (Silverman & Young, 1987).

Regarding the issues of the basic and smooth bootstrap, this present study presents a new multivariate kernel resampling technique for improving the performance of multiple regression with small samples because kernel technique has been proved remarkably successful for standard classification and regression problems (Schölkopf & Smola, 2002). KRT is a new resampling method where the overall procedure seems similar to that of the bootstrap, but KRT, by design, radically differs from the bootstrap in three-fold: (a) The bootstrap samples are randomly drawn from the exact original small sample data with replacement in the basic bootstrap, whereas the KRT samples are each randomly drawn from a *neighborhood* of a data point in the original small sample; (b) The kernel technique is used to select resamples with random noises instead of being used to smooth the resampling distributions as does the smooth bootstrap; and (c) KRT has a fixed bandwidth for the kernel used in the resampling procedures whereas the smooth bootstrap does not.

KRT utilizes Gaussian kernels (Silverman, 1986; Simonoff, 1996), the most commonly-used kernel technique (Yip, Ahmad, & Pong, 1999), to capture the underlying distribution of the given multivariate small sample data. The Gaussian kernel bandwidth is determined to produce an asymptotically optimal bandwidth minimizing the *mean integrated square error* (MISE; Silverman, 1986).

Specifically, let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be the given multivariate small sample data or n vectors from a d-dimensional space R^d, where d is the number of variables and n is the sample size. The KRT procedure is as follows:

Step 1. Define n multivariate Gaussian kernels as $K_i(\mathbf{x}) \sim N_d(\mathbf{X}_i, \mathbf{H}_0^2)$, $i = 1, \ldots, n$, where the mean vector $\mathbf{X}_i$ is the ith multivariate observation in the multivariate small sample and the random noise $\mathbf{H}_0$ can be determined by the optimal bandwidth MISE (Silverman, 1986, p. 87; Simonoff, 1996, p. 105):

$$\mathbf{H}_\text{o} = \left(\frac{4}{d+2}\right)^{1/(d+4)} \Sigma^{1/2} n^{-1/(d+4)}, \tag{1}$$

where Σ is the population covariance matrix and it can be estimated by $\mathbf{S}$, a sample covariance matrix of $\mathbf{X}_1, \ldots, \mathbf{X}_n$.

Step 2. Draw n multivariate random observations, $\mathbf{X}^*_1, \ldots, \mathbf{X}^*_n$, each from one multivariate Gaussian kernel $K_i(\mathbf{x})$ ($i = 1, \ldots, n$). The n multivariate random observations are defined as a multivariate KRT sample. According to Silverman (1986, eq. 4.7, p. 78) and Simonoff (1996, eq. 4.5, p. 102), the multivariate KRT sample has an estimated multivariate density function as follows:

$$\hat{f}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|}\sum_{i=1}^{n} k_d\left[\mathbf{H}^{-1}(\mathbf{x}-\mathbf{X}_i)\right], \tag{2}$$

where $k_d(\mathbf{u}) \sim N_d(\mathbf{0}, \mathbf{I})$ and has the following distributional relationship with the multivariate Gaussian kernel $K_i(\mathbf{x})$, $i = 1, \ldots, n$:

$$\begin{aligned} k_d(\mathbf{u}) &= (2\pi)^{-d/2}\exp\left(-\frac{1}{2}\mathbf{u}^\text{T}\mathbf{u}\right) \\ &= (2\pi)^{-d/2}\exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{X}_i)^\text{T}\mathbf{H}^{-2}(\mathbf{x}-\mathbf{X}_i)\right] \\ &= K_i(\mathbf{x}). \end{aligned} \tag{3}$$

Step 3. Conduct multiple regression over the multivariate KRT sample, $\mathbf{X}^*_1, \ldots, \mathbf{X}^*_n$, to obtain an estimate of a parameter of interest.

Step 4. Repeat Steps 2 to 3 k times, where k is called the KRT resampling parameter, to obtain k parameter estimates that comprise a sampling distribution of the parameter of interest.

Step 5. Evaluate the performance of KRT in regression analysis based on the sampling distribution such as estimation bias and standard errors.

The above procedure has been written into a SAS macro program and is available through the author. The "plug-in principle" (Efron & Tibshirani, 1998, p.35) makes the application of KRT very simple and straightforward because the above procedure does not require researchers to modify the bandwidth of the kernel to obtain KRT samples. Therefore, KRT is methodologically comparable to the bootstrap but the application of KRT is simpler than that of the bootstrap.

In the next section, an empirical example is presented for illustrating the application of the KRT procedure to multiple regression and evaluating the statistical performance of KRT in multiple regression with a small sample.

An Empirical Study of KRT in Multiple Regression Analysis

The Cement Hardening Data and Regression Model

The famous small sample of the Cement Hardening Data (CH) (Hjorth, 1994, p. 31) was used to study the performance of the application of the KRT procedure in multiple regression while comparing

Table 1. *The Cement Hardening Data*

x_1	x_2	x_3	x_4	y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

Note. x_1 = amount of tricalcium aluminate, $3C_aOAL_2O_3$; x_2 = amount of tricalcium silicate, $3CaOS_iO_2$; x_3 = amount of calcium aluminum ferrate, $4CaOAl_2O_3Fe_2O_3$; x_1 = amount of dicalcium silicate, $2CaOS_iO_2$; y (response) = heat evolved in calories per gram of cement.

The CH data (Table 1) with 13 observations depict the hardening of the cement and the heat evolved during the first 180 days after addition of water. x_1, x_2, x_3, and x_4 were linearly dependent predictors of the amounts of different components of chemicals, and the response variable y was the heat produced. Because Hjorth's (1994) linear regression model (4) was proven to have a good fit with inclusion of all the four predictors presented, it was used for both the KRT samples and the bootstrapping observations.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \text{, } i = 1, 2, ..., n. \tag{4}$$

Comparisons of Estimation Accuracy for the Empirical data

Table 2 shows the ordinary least-squares (OLS) estimates ($\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$) from the original small sample, the estimates from 200 KRT samples and 200 bootstrap samples and their estimated standard errors $\overline{se}\,(\hat{\beta}_i)$, i = 1, …, 4, from the regression model (4).

From Table 2 we can see that the estimates from KRT were systematically closer to the estimates from OLS than the estimates from the bootstrap; and the standard errors estimated from KRT were also systematically smaller than those from both the bootstrap and OLS. In addition, it is clear that the approximate biases of the KRT estimates were systematically smaller than the biases of the bootstrap estimates.

Table 3 shows the model fit indices from all the methods. We can see from Table 3 that the root mean squared error (RMSE) estimate from the KRT procedure was also close to the estimate from the OLS method, but the RMSE estimate from the bootstrap was still downward biased for the analysis on the empirical data. The multiple correlation R^2 from the KRT procedure is also comparable to that from both the bootstrap and OLS method (see Table 3).

Results and Discussions

The present study presents the new multivariate resampling method, KRT, for obtaining more accurate estimates with reasonable standard errors in multiple regression analysis with small samples. Unlike the smooth bootstrap, which draws resamples from the smoothed distribution of the bootstrap data, KRT draws the resamples from the neighborhoods of the original data with a fixed but optimal bandwidth. As such, of KRT has the following advantages: (a) the resample distribution strictly follows the original sample distribution, (b) the sampling distribution is not artificially modified, and (c) there is no need for researchers to consider the kernel bandwidth.

The findings from the applications of KRT in multiple regression to the empirical data suggest that the KRT procedure outperformed other methods in terms of the accuracy of the estimation of regression coefficients, estimation bias, and standard errors, comparing with the OLS method on the original small

Table 2. Estimates of Regression Analysis on the Cement Hardening Data

	OLS Small *N*		Bootstrap (*B* = 200)				KRT (*k* = 200)			
				Approx		Bias-Corrected		Approx		Bias-Corrected
Parameter	Est.	*SE*	Est.	Bias	*SE*	Est.	Est.	Bias	*SE*	Est.
β_0	62.41	70.07	74.00	11.59	117.10	50.81	61.10	-1.31	28.10	63.71
β_1	1.55	0.74	1.46	-0.09	1.16	1.64	1.57	0.02	0.30	1.53
β_2	0.51	0.72	0.39	-0.12	1.22	0.63	0.52	0.01	0.29	0.50
β_3	0.10	0.75	0.01	-0.09	1.20	0.20	0.12	0.02	0.30	0.08
β_4	-0.14	0.71	-0.27	-0.13	1.21	-0.02	-0.13	0.01	0.28	-0.16

Note. Est. = Estimate. Approx = Approximate.

Table 3. Model Fit Summary for the Cement Hardening Data

	OLS	Bootstrap (*B* = 200)			KRT (*k* = 200)		
Parameter	Small *N*	Est.	Approx Bias	*SE*	Est.	Approx Bias	*SE*
RMSE	2.45	1.84	-0.60	0.55	2.59	0.15	0.28
R^2	0.98	0.99	0.01	0.01	0.98	0.00	0.01

Note. Est. = Estimate. Approx = Approximate.

sample and the bootstrap results. The results from this study also support the use of KRT as a viable alternative to improving the performance of multiple regression analysis with small samples. This current study suggests that KRT can be a useful tool for researchers to conduct multiple regression analysis when only small samples are available. It will help researchers draw more valid statistical inference than using the original small sample.

The advantage of KRT concerns the resampling procedure's simplicity and efficiency. Compared to the bootstrap, KRT obtains comparable or more accurate estimates, but does not require researchers to modify the complicated resampling procedures. The results from this study indicated that the KRT procedure has overcome the two major limitations that the bootstrap method encountered. First, KRT obtains independent resamples through sampling from the neighborhoods of the data points, which solves the lack of independent observations of the basic bootstrap resamples. Secondly, the KRT procedure practically advances the smooth bootstrap by using fixed optimal bandwidth for the kernel procedure instead of requiring researchers to customize the optimal bandwidth to their data. The simplicity of KRT will help improve the practical applications of resampling method in the real research and promote the use of resampling method in the computer age.

In the present study, we only explored the statistical performance of the application of the KRT procedure to the multiple regression in terms of the estimation of regression coefficients and model fit indices. Even though we have used the population parameters to verify the smaller biases for the estimations from the KRT procedure for the both estimates and standard errors, significance tests and the confidence intervals are desirable for further research.

References

Aizerman, M., Braverman, E., & Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control,* 25, 821-837.

Allision, P. D. (1999). *Multiple regression: A primer*. Thousand Oaks, CA: Pine Forge Press.

Bai, H., & Pan, W. (2008). Resampling methods revisited: Advancing the understanding and applications in educational research. *International Journal of Research & Method in Education, 31*(1), 45-62.

Bickel, P. J., & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, *9*, 1196-1217.

Bickel, P. J., & Freedman, D. A. (1983). Bootstrapping regression models with many parameters. In P. J. Bickel, K. Doksum, & J. L. Hodges (Eds.), *Festschrift for Erich Lehmann* (pp. 28-48). Belmont, CA: Wadsworth.

Chernick, M. R. (1999). *Bootstrap methods: A practitioner's guide*. New York: Wiley and Sons, Inc.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*. New York: Cambridge University Press.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1-26.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics* (No. 38). Philadelphia: SIAM.

Efron, B., & Tibshirani, R. J. (1998). *An introduction to the bootstrap*. New York: Chapman & Hall/CRC.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics*, *9*(6), 1218-1228.

Hjorth, J. S. U. (1994). *Computer intensive statistical methods*. London: Chapman & Hall.

Peters, S. C., & Freedman, D. A. (1984). Some notes on the bootstrap in regression problems. *Journal of Business and Economic Statistics, 2*, 401-409.

SAS Institute Inc. (2008). *Sample 24982: Jackknife and bootstrap analyses.* Retrieved May 2, 2008, from http://support.sas.com/kb/24/982.html

Schölkopf, B., & Smola, A. J. (2002). *A short introduction to learning with kernels*. New York: Springer-Verlag.

Shao, J. (1988). On resampling methods for variance and bias estimation in linear models. *The Annals of Statistics, 16*, 986-1008.

Shimabukuro, F. I., Lazar, S., Dyson, H. B., & Chernick, M. R. (1984). A quasi-optical method for measuring the complex permittivity of materials. *IEEE Transactions on Microwave Theory and Techniques, 32*, 659-665.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Silverman, B. W., & Young, G. A. (1987). The bootstrap: To smooth or not to smooth? *Biometrika, 74*, 469–79.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. New York: Springer.

Towers, S. (2002). Kernel probability density estimation methods. *Proceedings on Advanced Statistical Techniques in Particle Physics*, Durham, England.

Weber, N. C. (1984). On resampling techniques for regression models. *Statistics & Probability Letters, 2*, 275-278.

Wu, C. F. J. (1986). Jackknife bootstrap and other resampling plans in regression analysis (with discussion). *The Annals of Statistics, 14*, 1261-1350.

Yip, H. M., Ahmad, I., & Pong, T. C. (1999). An efficient parallel algorithm for computing the Gaussian convolution of multi-dimensional image data. *Journal of Supercomputing*, *14*, 233-255..

Send correspondence to: Haiyan Bai
University of Central Florida
Email: hbai@mail.ucf.edu