

# Assessing the regression to the mean for non-normal populations via kernel estimators

Majnu John<sup>1</sup>, Abbas F. Jawad<sup>2</sup>

<sup>1</sup>Division of Biostatistics and Epidemiology, Department of Public Health, Weill Cornell Medical College, New York, USA.

<sup>2</sup>Division of Biostatistics and Epidemiology, Department of Pediatrics, University of Pennsylvania, Philadelphia, PA, USA.

**Citation:** John M, Jawad AF. Assessing the regression to the mean for non-normal populations via kernel estimators. *North Am J Med Sci* 2010; 2: 288-292.

**Availability:** www.najms.org

**ISSN:** 1947 – 2714

## Abstract

**Background:** Part of the change over time of a response in longitudinal studies may be attributed to the regression to the mean. The component of change due to regression to the mean is more pronounced in the subjects with extreme initial values. Das and Mulder proposed a nonparametric approach to estimate the regression to the mean. **Aim:** In this paper, Das and Mulder's method is made data-adaptive for empirical distributions via kernel estimation approaches, while retaining the original assumptions made by them. **Results:** We use the best approaches for kernel density and hazard function estimation in our methods. This makes our approach extremely user friendly for a practitioner via the state of the art procedures and packages available in statistical softwares such as SAS and R for kernel density and hazard function estimation. We also estimate the standard error of our estimates of regression to the mean via nonparametric bootstrap methods. Finally, our methods are illustrated by analyzing the percent predicted FEV1 measurements available from the Cystic Fibrosis Foundation's National Patient Registry. **Conclusion:** The kernel based approach presented in this article is a user-friendly method to assess the regression to the mean in non-normal populations.

**Keywords:** regression to the mean, kernel density estimation, kernel estimators for hazard function, bootstrap methods, longitudinal clinical studies

**Correspondence to:** Majnu John, Division of Biostatistics and Epidemiology, Department of Public Health, Weill Cornell Medical College, 402 East 67<sup>th</sup> Street, New York, NY 10065, USA. Email: maj2023@med.cornell.edu

## Introduction

In longitudinal clinical studies, where the change over time of an outcome is of interest, it is sometimes seen that the change is dependent on the initial value. The change over time may be significantly larger in subjects who had really higher (or lower) initial values as compared to subjects whose initial values were closer to the population mean. This differential effect may partly be due to a phenomenon known as *regression to the mean*. In other words, since there is always within-subject variability, the initial value may have been high (or low) for certain subjects just by random chance, and upon remeasurement the outcome values for these subjects would have just 'regressed' back to the true mean.

For example, this phenomenon may be illustrated via the analysis results of data in a registry based on cystic

fibrosis (CF) patients that was conducted by investigators at Children's Hospital of Philadelphia [1]. One of the goals of the study was to determine the rate of change of the pulmonary function as measured by percent of predicted forced expiratory volume in 1 second (FEV<sub>1</sub>%), over a 4-year period in a large cohort of children with cystic fibrosis. CF foundation's National CF Patient Registry data collected from 1991 to 1995 for 968 children (461 male) aged 5 to 8 years with pancreatic insufficiency and FEV<sub>1</sub>% between 60% and 140% were analyzed longitudinally. The significant decline in FEV<sub>1</sub>% was found to be dependent on baseline FEV<sub>1</sub>%; children with initial FEV<sub>1</sub>%  $\geq 90$  declined 2.6 U/y more than those with initial FEV<sub>1</sub>%  $< 90$ .

Various methods have been proposed in the literature to account for regression to the mean. These approaches

may be broadly classified into two categories: 1) methods where the regression to the mean is estimated and then subtracted from or added to the observed change to estimate the true change; 2) methods which check for the relationship between change and initial value via correlation or regression, where the correlation co-efficient or the regression co-efficient as the case may be adjusted for the regression towards mean. Methods of the first type have been proposed and studied in [2-8]. Methods of the second type have been proposed in [9-10]. In this paper we propose a method of the first type. We present an estimator for the regression to mean when the underlying distribution for the measurements are non-normal.

Gardener, James and Davis [2-4] discussed estimators for regression to mean when the measurements are normally distributed with mean and variance that remain constant over time. Chinn and Heller [5] presented similar results when the mean and variance were assumed to vary over time. Das and Mulder [6] investigated the effect when the measurements were not necessarily normal, however their method is not directly applicable to empirical distributions. Their method was made more usable for a practitioner/analyst by a method proposed in Sheath and Dobson [7] where the regression to mean was estimated for empirical distributions using Edgeworth series and saddle point approximations. Muller, Abramson and Azari [8] propose an elegant nonparametric method for estimating the regression to the mean under fewer assumptions related to the underlying distributions. In this paper, Das and Mulder's method [6] is made data-adaptive for empirical distributions via kernel estimation approaches, while retaining the original assumptions made by them.

We adopted the best approaches for kernel density and hazard function estimation in our methods. This makes our approach extremely user friendly for a practitioner via the state of the art procedures and packages available in statistical softwares such as SAS and R for kernel density and hazard function estimation.

### An estimate for regression to the mean in non-normal populations

In this paper we restrict our attention to the studies which measure only one additional value other than the initial value. Following the notational convention in [7], let  $X_{1i}$  and  $X_{2i}$  be random variables representing the initial and follow-up measurements on the  $i^{th}$  subject,  $i = 1, \dots, n$ , where  $X_{1i} = M_i + e_{1i}$  and  $X_{2i} = M_i + e_{2i}$ .

Here  $M_i$ ,  $e_{1i}$ , and  $e_{2i}$  are mutually independent random variables representing the true value, measurement error or within-subject variability for the initial and follow-up measurements respectively. We

also assume that the  $M_i$ 's are i.i.d. with mean  $\mu$  and variance  $\theta^2 = \rho\sigma^2$  and  $e_1$  and  $e_2$  have the normal distributions  $N(0, \Delta^2)$ , where  $\Delta^2 = (1 - \rho)\sigma^2$ .

For the initial measurements  $X_{1i}$ 's let  $x_L$  denote the cut-off point for 'large' values; that is all the subjects with initial measurements greater than  $x_L$  are considered subjects with high initial value. We further assume that all the  $X_{1i}$ 's are i.i.d. with a probability density function  $g(x)$  and distribution function  $G(x)$ .

Then the expression for evaluating the regression to the mean for the subjects with high initial values, as proposed by Das and Mulder [6] is,

$$R(x_L) = E((X_1 - X_2) | X_1 > x_L) = \frac{(1 - \rho)\sigma^2 g(x_L)}{1 - G(x_L)} \tag{1}$$

We propose to estimate  $R(x_L)$  by two different approaches as outlined in sections 3 and 4. Let  $g$  be a kernel density estimator for  $g$  and let  $\hat{G}$  be the empirical distribution function for  $X_{1i}$ 's greater than  $x_L$ . Then in section 3, we propose an estimator  $\hat{R}(x_L)$  of  $R(x_L)$  given by

$$\hat{R}(x_L) = \frac{(1 - \rho)\sigma^2 \hat{g}(x_L)}{1 - \hat{G}(x_L)} \tag{2}$$

Let us denote  $u(x) = \frac{g(x)}{1 - G(x)}$ . When  $X_1$  is a non-negative

variable,  $u(x)$  is similar to the hazard rate function used in survival analysis literature. We may find an estimator  $\hat{u}(x)$  of  $u(x)$  via the state of the art kernel estimation techniques for hazard rate function. This approach is outlined in section Estimation of regression to the mean via a kernel estimator for  $u$ .

### Estimation of regression to the mean via a kernel estimator for $g$

Kernel density estimates have been studied extensively in literature and used in various applications, ever since they were proposed in [11, 12]. For the initial values,  $X_{1i}$ ,  $i = 1, \dots, n$ , the kernel density estimator of the probability density function  $g$  is given by

$$g_h(x) = n^{-1} \sum_{i=1}^n K_h(x - X_{1i})$$

where  $K_h(\cdot) = \frac{1}{h} K(\cdot/h)$  for a 'kernel function'  $K$

(often taken to a symmetric probability density) and a 'bandwidth'  $h$  (the smoothing parameter). (See [13] for a good introduction to the ideas.) A common way of measuring the estimation error of  $\hat{g}_h(x)$  is via the mean

integrated squared error (MISE),

$$MISE(h) = E \int (\hat{g}_h - g)^2$$

where  $\int$  denotes the integration over the real line.

Asymptotic mean integrated square error (AMISE) given by

$$AMISE(h) = n^{-1} h^{-1} R(K) + h^4 R(f'') \left( \int x^2 K^2 \right)^2,$$

is a more useful criterion which provides simple insight into "good" bandwidths. Here  $R(\varphi) = \int \varphi^2(x) dx$  and  $\int x^2 K = \int x^2 K(x) dx$ . The minimizer of AMISE ( $h$ ) given by

$$h_{AMISE} = \left[ \frac{R(K)}{nR(f'') \left( \int x^2 K \right)^2} \right]^{1/5}$$

is a good approximation to  $h_{MISE}$ , the minimizer of MISE, in many circumstances.

Various methods that have been proposed for optimal bandwidth selection may be broadly classified as *first generation* and *second generation methods* [14]. The most popular among the first generation methods are *rules of thumb* [13], *least squares cross-validation* [15-17], and *biased cross-validation* [18]. Second generation methods include *solve-the-equation plug-in approach* [19,20] and *smoothed bootstrap* [21-23]. As shown in [14], the method proposed in [20] is a stable and consistent method that outperforms all the other methods mentioned above. The idea behind this method is to choose the bandwidth that is a solution of the fixed point equation

$$h = \left[ \frac{R(K)}{nR(f''(h)) \left( \int x^2 K \right)^2} \right]^{1/5}$$

This is the bandwidth selection method that we use for the kernel density estimate  $\hat{g}(h)$ . This is also the default method in many statistical software packages including PROC KDE in SAS version 9.0.

## Estimation of regression to the mean via a kernel estimator for $u$

For non-negative  $X_{1i}$ , then the function  $u$  defined in section 2 is become the hazard function, if we interchange the 'time to event' variable with  $X_{1i}$  and assume no 'censoring'. Kernel based estimation methods of hazard function have received considerable attention in the statistical literature.

Kernel estimators for the hazard function was initially proposed in [24] and was further studied in [25]. [26] gave a generalized form to these estimators and studied the asymptotic properties of the generalized estimator.

Adapting the generalized form to our notation for regression to mean estimator, we have

$$\hat{u}_h(x) = n^{-1} \sum_{i=1}^n K_h \left( \frac{x - X_{1(i)}}{h} \right) \frac{1}{n - i + 1}$$

as a kernel estimator for  $u$ . Here  $X_{1(i)}$ 's denote the ordered sample of  $X_{1i}$ 's.

As in the case of kernel density estimation, an important problem when estimating the hazard function by kernel methods is the choice of the smoothing parameter (that is, the bandwidth)  $h$ . Various methods have been proposed in the statistical literature for choosing the optimal bandwidth. A maximum likelihood cross-validation method for uncensored hazard estimation was proposed in [27]. Later, [28] (via simulation results) and [29] (via theoretical insights) have shown that least squares cross validation as a more appropriate approach. A least squares cross-validation methods proposed in [30] for density estimation was extended to uncensored hazard function estimation in [31] and in [32]. Later [33] extended these methods via a bootstrap selection of the smoothing parameter.

All the methods mentioned above are fixed-bandwidth fixed-kernel methods for estimating the hazard function. When the data is not evenly distributed over the range of interest, the degree of smoothing achieved via a fixed-bandwidth method will not be uniform. Using a varying bandwidth estimator which balances the local variance and local bias as suggested in [34] rectifies the non-adaptive behavior of fixed-bandwidth estimators. Another approach is to incorporate the idea of 'nearest neighbor' into the definition of bandwidth [35]. Fixed kernel estimators do not take into account the so-called *boundary effects* (that is, the bias problems) near the endpoints of the support of the hazard function. One way to overcome this problem is to change the kernels at the boundary [36-38]. [34] presented a class of boundary kernels which gave rise to smaller leading constants of the asymptotic mean squared error when estimating than other boundary kernels considered previously. [39] confirmed the advantages of methods proposed in [34] and [35] over the previously published methods for estimating hazard function via kernel estimators. These two methods are available in a R package named *mu haz* downloadable from [cran.r-project.org](http://cran.r-project.org).

## Bootstrap estimates of standard error

Bootstrap techniques have gained tremendous popularity ever since it was introduced in the seminal paper [40]. Bootstrap methods are well known for estimating the standard errors and bias. A good introduction to the bootstrap world can be seen in [41] and in [42].

Let us denote by  $\hat{R}$ , our estimate of the regression to the mean via any of the methods proposed in sections 3 and 4. We may obtain the bootstrap estimate of standard error and bias via the following steps. First we obtain  $B$  nonparametric bootstrap samples of the data. A nonparametric bootstrap sample is obtained by sampling with replacement from the empirical distribution of the original data. For each of the  $B$  bootstrap samples, we calculate our estimate of regression to the mean, denoted by  $\hat{R}_1^*, \dots, \hat{R}_B^*$ . The bootstrap estimate of the standard error of  $\hat{R}$  is simply the standard deviation of  $\hat{R}_1^*, \dots, \hat{R}_B^*$ . That is, the standard error of  $\hat{R}$  is obtained as

$$se_B(\hat{R}) = \sqrt{\frac{1}{B} \sum_{i=1}^B (\hat{R}_i^* - \bar{R}^*)^2}$$

where

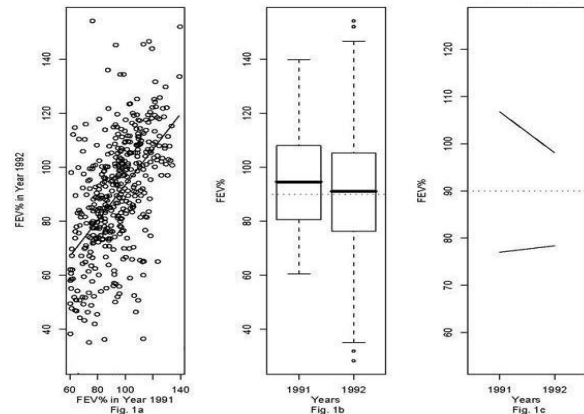
$$\bar{R}^* = \frac{1}{B} \sum \hat{R}_i^*$$

## Application to percent predicted FEV1 measurements in cystic fibrosis patients

We illustrate the methods described in this paper using percent predicted FEV 1 measurements of 461 males in the CF registry obtained in the years 1991 and 1992. The correlation between the measurements from years 1991 and 1992 was found to be 0.55 and variance of the 1991 measurements was 323.3. (See Figures 1a and 1b.) The PROC MIXED procedure in SAS with a random intercept was used separately for subjects with the initial value (that is, 1991 measurement) greater than and less than or equal to 90. For the subjects with initial value above 90, the intercept was 106.76 and the slope was -8.69 with standard errors 0.97 and 1.38 respectively and for the subjects with initial value below or equal to 90, the intercept was 77.00 and the slope was 1.37 with standard errors 1.08 and 1.54 respectively. (See Figure 1c.) We suspect that the regression to the mean plays a significant role in the change seen in subjects with high initial value.

We assess the regression to the mean in subjects with initial measurements greater than 90 via the two methods described in this paper. For the method in the section 3, we used the Gaussian kernel along with the plug-in bandwidth selection method described in [20] to estimate  $g$ . This was done via PROC KDE in SAS. The regression to the mean via this method was estimated to be -4.78 with bootstrap standard error of 0.85. Since the percent predicted FEV1 measurements is a non-negative variable, we were also able to estimate the regression to the mean via the method described in section 4. We used the varying kernel, varying bandwidth method suggested in [34] with the Epanechnikov boundary kernel to estimate  $u$ . This was done using the *muhaz* R package. (See section 4 for more

details.) The regression to mean estimate via this method was seen to be -5.21 with bootstrap standard error of 1.02. The regression to the mean estimate by both the methods are comparable in this case. Via the first method we see that 55% of the change in percent predicted FEV1 between 1991 and 1992 is due to the regression to the mean effect and by the second method we see that 60% of the change is attributable to regression to the mean.



**Fig. 1** Plots. **1a:** Scatter plot of FEV% of male subjects at years 1991 and 1992, **1b:** Box plots of FEV% in years 1991 and 1992. **1c:** Linear mixed effects model fit for subjects with initial measurement  $> 90$  and for subjects with initial measurement  $\leq 90$ .

## Concluding remarks

We propose two approaches for estimating Das and Mulder's expression of regression to the mean for non-normal population. Our first approach is based on a kernel density estimation method and our second approach is based on kernel estimation approaches for hazard rate function. The estimates in both the methods may be easily obtained via the state of the art techniques available in popular statistical software such as SAS and R. We also calculate the bootstrap standard error for estimates. Applying our methods to FEV % data from Cystic Fibrosis foundation's National Patient Registry showed that both the regression to the mean estimates are within the margin of bootstrap standard error estimates. As in Das and Mulder's original paper, our methods are restricted to calculating the regression to the mean when there are only two time points. Extension of these methods to longitudinal studies with more than two time points is an area of future research.

## Acknowledgements

Partial support for the first author came from Clinical Translational Science Center grant (UL1-RR024996).

## References

1. Zemel BS, Jawad AF, Fitz Simmons S, Stallings YA. Longitudinal relationship among growth, nutritional status, and pulmonary function in children with cystic fibrosis: Analysis of the Cystic Fibrosis Foundation National CF Patient Registry. *J Pediatr* 2000; 37 (3)

- 374-380.
2. Gardner MJ, Heady JA. Some effects of within-person variability in epidemiological studies. *J Chronic Dis* 1973; 26, 781-795.
  3. James KE. Regression towards mean in uncontrolled clinical trials. *Biometrics* 1973; 29, 121-130.
  4. Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. *Am J Epidemiol* 1976; 104, 493-498.
  5. Chinn S, Heller RF. Some further results concerning regression to the mean. *Am J Epidemiol* 1981, 114, 902-905.
  6. Das P, Mulder PGH. Regression to the mode. *Stat Neerl* 1983; 37, 15-20.
  7. Beath KJ, Dobson, AJ. Regression to the mean for nonnormal populations. *Biometrika* 1991; 78, 431-435.
  8. Muller H, Abramson I, Azari R. Nonparametric regression to the mean. *PNAS* 2003; 100 (17), 9715-9720.
  9. Blomqvist N. On the relation between change and initial value. *JASA* 1977; 72, 746-749.
  10. Oldham PD. A note on the analysis of repeated measurements of the same subjects. *J Chronic Dis* 1962; 15, 969-977.
  11. Rosenblatt M. Remarks on some nonparametric estimates of a density function. *Ann Math Stat* 1956; 27, 832-837.
  12. Parzen E. On estimation of a probability density function and mode. *Ann Math Stat* 1962; 33 (3), 1065-1076.
  13. Silverman BW: *Density estimation for statistics and data analysis*. London: Chapman and Hall; 1986
  14. Jones MC, Marron JS, Sheather SJ. A brief survey of bandwidth selection for density estimation. *JASA* 1996; 91, 401-407.
  15. Bowman AW. An alternative method for the smoothing of density estimates. *Biometrika* 1984; 71, 353-360.
  16. Rudemo M. Empirical choice of histograms and kernel density estimators. *Scand J Stat* 1982; 9, 65-78.
  17. Hall P, Marron JS. Local minima in cross-validation functions. *J R Stat Soc Ser B* 1991; 53, 245-252.
  18. Scott DW, Terrell CR. Biased and unbiased cross-validation in density estimation. *JASA* 1987; 82, 1131-1146.
  19. Hall P. Objective methods for the estimation of window size in the nonparametric estimation of a density. The Australian National University, Mathematical Science Institute, Technical Report 1980.
  20. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc, Ser B* 1991; 53, 683-690.
  21. Faraway JJ, Jhun M. Bootstrap choice of bandwidth for density estimation. *JASA* 1990; 85, 1119-1122.
  22. Taylor CC. An alternative method for the smoothing of density estimates. *Biometrika* 1984; 76, 705-712.
  23. Marron JS: Bootstrap bandwidth selection. In Le Page R, Billiard L. eds. *Exploring the limits of bootstrap*, New York, NY: John Wiley; 1992:249-262.
  24. Watson CS, Leadbetter MR. Hazard Analysis 1. *Biometrika* 1964; 51, 175-184.
  25. Nelson W. Theory and application of hazard plotting for censored failure data. *Technometrics* 1972; 14, 945-965.
  26. Ramlau-Hansen, H. Smoothing counting process intensities by means of kernel functions. *Ann Stat* 1983; 11, 453-466.
  27. Tanner MA, Wong WH. Data based nonparametric estimation of the hazard function with applications to model diagnostics and exploratory analysis. *JASA* 1984; 79, 174-182.
  28. Cao R, Cuevas A, Gonzalez-Manteiga W. A comparative study of several smoothing methods in density estimation. *Comput Statist Data Anal* 1994; 17, 153-176.
  29. Hall P. On Kullback-Leibler loss and density estimation. *Ann Stat* 1987; 15, 1491-1519.
  30. Marron JS, Padgett JW. Asymptotically optimal bandwidth selection for a kernel density estimator. *Ann Stat* 1987; 15, 1520-1535.
  31. Patil PN. Bandwidth choice for nonparametric hazard rate estimation. *J Stat Plan Inference* 1993; 35, 15-30.
  32. Sarda P, Vieu P. Smoothing parameter selection in hazard estimation. *Stat Probabil Letters* 1991; 11, 429-434.
  33. Gonzalez-Manteiga W, Cao R, Marron JS. Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *JASA* 1996; 91, 1130-1140.
  34. Muller HG, Wang JL. Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* 1994; 50, 61-76.
  35. Gefeller O, Dette H. Nearest neighbor kernel estimation of the hazard function from censored data. *J Statist Comput Simul* 1992; 43, 93-101.
  36. Hougaard P. A boundary modification of kernel function smoothing, with application to insulin absorption kinetics. *Compstat* 1988; 31-36. Berlin: Springer.
  37. Hougaard P, Plum A, Ribel U. Kernel function smoothing of insulin absorption kinetics. *Biometrics* 1989; 45, 1041-1052.
  38. Hall P, Wehrly TE. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *JASA* 1991; 86, 665-672.
  39. Hess KR, Serachitopol DM, Brown BW. Hazard function estimators: a Simulation study. *Stat Med* 1999; 18, 3075-3088.
  40. Efron B. Bootstrap methods: Another look at the jackknife. *Ann Stat* 1979; 7, 1-26.
  41. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*, New York: Chapman and Hall, 1993.
  42. Boos DD. Introduction to the Bootstrap world. *Stat Sci* 2003; 18 (2), 168-174.