Clinical Chemistry 55:1
165–169 (2009)

Brief Communications

Using Median Regression to Obtain Adjusted Estimates of Central Tendency for Skewed Laboratory and Epidemiologic Data

Katharine M. McGreevy,[1] Stuart R. Lipsitz,[2*] Jeffrey A. Linder,[2] Eric Rimm,[3] and David G. Hoel[4]

[1] New Jersey Department of Health and Senior Services, Trenton, NJ; [2] Division of General Medicine, Harvard Medical School, Boston, MA; [3] Departments of Epidemiology and Nutrition, Harvard School of Public Health, Boston, MA; [4] Department of Biostatistics, Bioinformatics, and Epidemiology, Medical University of South, Carolina, Charleston, SC; * address correspondence to this author at: Division of General Internal Medicine, Brigham and Women's Hospital, 1 Brigham Circle, 1620 Tremont St., 3rd Floor BC3 002D, Boston, MA 02120. E-mail slipsitz@partners.org.

BACKGROUND: Laboratory studies often involve analyses of highly skewed data for which means are not an adequate measure of central tendency because they are sensitive to outliers. Attempts to transform skewed data to symmetry are not always successful, and medians are better measures of central tendency for such skewed distributions. When medians are compared across groups, confounding can be an issue, so there is a need for adjusted medians.

METHODS: We illustrate the use of quantile regression to obtain adjusted medians. The method is illustrated by use of skewed nutrient data obtained from black and white men attending a prostate cancer screening. For 3 nutrients, saturated fats, caffeine, and vitamin K, we obtained medians adjusted by age, body mass index, and calories for men in each race group.

RESULTS: Quantile regression, linear regression, and log-normal regression produced substantially different adjusted estimates of central tendency for saturated fats, caffeine, and vitamin K.

CONCLUSIONS: Our method was useful for analysis of skewed and other nonnormally distributed continuous outcome data and for calculation of adjusted medians.

Many types of laboratory and epidemiologic studies involve the analyses of highly skewed data. When the distributions of outcome variables are highly skewed, the mean is sensitive to outliers and is not a good measure of central tendency. Furthermore, when a skewed outcome variable is compared across groups, use of a Student *t*-test to compare group means or an ordinary least-squares regression to compare adjusted group means can yield biased results, including loss of statistical power to detect a true difference or a high type I (false-positive) error rate when there is no difference. A transformation of the outcome variable is a popular approach to improve symmetry and normality for linear regression. Because transformations make interpretations difficult and often fail to normalize the data, however, the median is often reported as a measure of central tendency. When medians are compared across groups, there is frequently a need to adjust for potential confounders such as age. We propose the use of adjusted medians instead of adjusted means.

We obtained the adjusted medians from quantile regression *(1)*. In statistics quantile is synonymous with percentile (although strictly speaking, a quantile is a fraction), and quantile regression is a general statistical technique to estimate the effect of covariates on any specified percentile, not just the median (50th percentile). For instance, we could estimate adjusted 25th percentiles in different groups by using quantile regression and then compare these adjusted 25th percentiles across groups. When used to estimate the median, quantile regression is usually called median regression. Median regression is also formally known as least absolute-value (LAV) regression, because the median regression parameters are estimated by minimizing the sum of the absolute value of the residuals. In contrast, the linear regression parameters are estimated by minimizing the sum of the squares of the residuals, hence the name ordinary least squares regression. Appendix 1 (available electronically with the online version of this paper at www.clinchem.org) contains a brief review of the LAV estimation technique we used to obtain the estimated adjusted medians. If there are no covariates in the median regression model (only an intercept), then the estimated intercept is just the usual estimate of the median (middle number in the dataset). The adjusted median obtained with LAV is less sensitive to outliers than an adjusted mean, analogous to an unadjusted median vs an unadjusted mean. Although we found recent reports of the use of quantile regression *(2–8)*, it is an underused technique.

We used adjusted medians obtained from quantile (median) regression to compare nutrient intakes between black and white males. The nutrient data were obtained by using the Gladys Block Brief 2000 dietary questionnaire *(2)*, which was completed by 138 white and 95 black men who attended a free annual prostate cancer screening held at the Medical University of South Carolina in 2002 *(9, 10)*. This study was approved by the Institutional Review Board of the Medical University of South Carolina, and informed consent was obtained from all participants. It is well

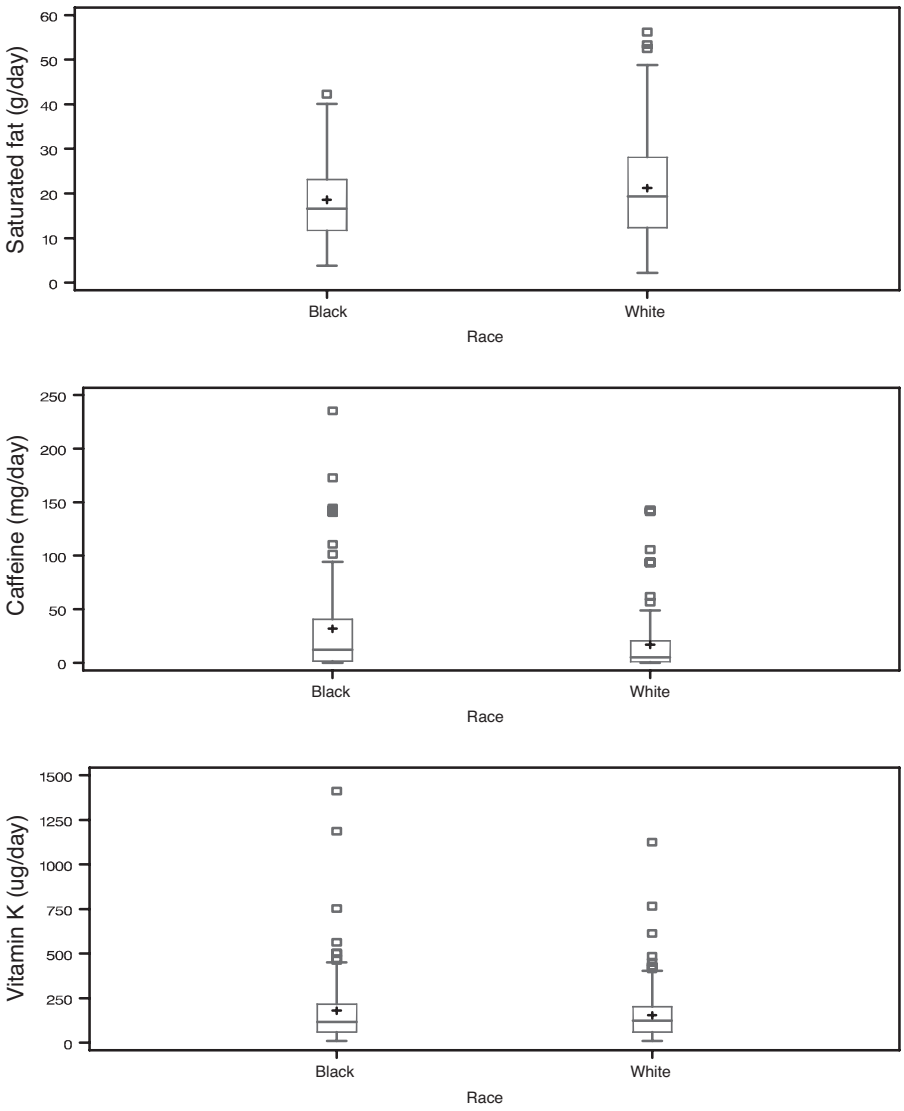

Fig. 1. Box-plots for saturated fats, caffeine, and vitamin K, by race group.

established that black males have an increased risk of developing prostate cancer compared to white males *(11)*, and it is possible that some of the racial differences in prostate cancer risk may be related to differences in nutritional intake. This study explored the nutritional intake of black and white males.

Data regarding the daily intake of 36 nutrients were obtained from the questionnaire, and the data distributions for all 36 nutrients were found to be asymmetric by the Mardia test at a Bonferroni-corrected significance level of 5%/36 = 0.14%. We present data for 3 of these nutrients, saturated fats, caffeine, and vitamin K. In box plots by race for these 3 variables, distributions for all variables were skewed to the right, e.g., the mean was greater than the median (Fig. 1). Although we rejected the null hypothesis of symmetry for each variable, in the dataset the saturated fats variable was the least skewed (skewness = 0.91), caffeine was intermediate with regard to skewness (skewness = 2.5), and vitamin K had the most skewness (skewness = 3.5). For reference, a symmetric distribution has skewness = 0, and a distribution that is skewed to the right has positive skewness. The log transformation appeared to symmetrize vitamin K (Mardia $P > 0.6$), but not the other 2 variables ($P < 0.006$).

When comparing medians of the nutrients across races, we adjusted for the a priori potential confound-

Table 1. Estimated median nutrients per day by race.[a]

Nutrient	Whites		Blacks		*P*-value for equal medians
	Median	95% CI	Median	95% CI	
Saturated fat, g/day					
Unadjusted mean[b]	21.2	(19.3, 23.1)	18.6	(16.8, 20.4)	0.060
Unadjusted median[c]	19.2	(16.7, 21.7)	16.6	(14.4, 18.8)	0.106
Linear regression	21.3	(20.2, 22.3)	19.0	(17.9, 20.1)	0.003
Log-normal regression	18.4	(17.3, 19.7)	17.1	(16.0, 18.3)	0.020
Quantile regression	21.4	(19.8, 23.0)	18.4	(16.7, 20.2)	0.021
Caffeine, mg/day					
Unadjusted mean[b]	32.8	(23.5, 42.2)	18.0	(13.1, 23.0)	0.003
Unadjusted median[c]	13.4	(4.5, 22.2)	6.7	(3.9, 9.5)	0.026
Linear regression	31.3	(21.4, 41.2)	18.1	(13.0, 23.3)	0.021
Log-normal regression	7.2	(4.6, 11.2)	5.3	(3.8, 7.3)	0.267
Quantile regression	15.3	(4.4, 26.1)	8.4	(5.7, 11.1)	0.229
Vitamin K, μg/day					
Unadjusted mean[b]	180.6	(137.1, 224.0)	154.3	(129.5, 179.2)	0.273
Unadjusted median[c]	117.6	(88.3, 146.9)	123.5	(101.2, 145.7)	0.801
Linear regression	199.4	(152.6, 246.2)	147.2	(122.9, 171.5)	0.053
Log-normal regression	124.6	(102.3, 151.8)	103.2	(90.1, 118.2)	0.123
Quantile regression	130.3	(98.4, 162.1)	105.7	(87.5, 123.8)	0.188

[a] Adjusted for age, body mass index, and total calories and unadjusted.
[b] The 2-sample *t*-test is used to test for equal means when unadjusted means are used.
[c] The Wilcoxon rank-sum test is used to test for equal medians when the unadjusted median is used.

ers age and body mass index. In addition, we followed the convention *(13)* of adjusting for total calories to remove systematic over- or underreporting of dietary intake. We used the new SAS procedure Proc Quantreg (downloaded from the SAS Web site http://support.sas.com/rnd/app/da/quantreg.html for SAS 9.1) to obtain the adjusted medians and their SEs. Appendix 2 of the SAS documentation gives a sample of SAS coding for median regression; also, an SAS macro to calculate the estimates can be obtained from the first author. Alternatively, most statistical packages can calculate adjusted quantiles; for example, Stata function qreg and R (or S-plus) function quantreg. The estimated adjusted medians in blacks was

$$\hat{M}_B = \hat{\beta}_{0B} + \hat{\beta}_{1B}\,\overline{age} + \hat{\beta}_{2B}\,\overline{BMI} + \hat{\beta}_{3B}\,\overline{calories}$$

and in whites was

$$\hat{M}_W = \hat{\beta}_{0W} + \hat{\beta}_{1W}\,\overline{age} + \hat{\beta}_{2W}\,\overline{BMI} + \hat{\beta}_{3W}\,\overline{calories},$$

where the $\hat{\beta}$'s were estimated regression coefficients from the median regression (see Appendix 1 of the SAS documentation), and $\overline{age} = 58.6$, $\overline{BMI} = 27.0$, and $\overline{calories} = 1621$ were the sample means, pooled for blacks and whites in the study. We then tested for differences in the adjusted medians between blacks and whites by using a 2-sample Z-test (Appendix 1 of the SAS documentation), with the numerator being equal to $(\hat{M}_B - \hat{M}_W)$.

Adjusted medians and 95% CIs, as well as the *P*-values for testing equal adjusted medians are given in Table 1. We also give adjusted means from ordinary least squares linear regression with covariates race, age, body mass index, and total calories. These adjusted means could be considered to be adjusted medians if the nutrients were normally distributed, because the mean equals the median. We also give adjusted medians from log-normal linear regression, in which a log-transformed nutrient was the outcome for the linear regression, and then adjusted medians were obtained by exponentiating the adjusted log-normal means. Finally, we give unadjusted means and medians.

For the least-skewed variable (saturated fat), the adjusted estimates from linear regression and quantile regression were similar (Table 1), as might be expected for mildly skewed data. The adjusted estimates from log-normal regression were a few grams lower than those from quantile regression, a finding that suggests that the log-transformation might have skewed the dis-

tribution slightly to the left. The adjusted (from linear regression) and unadjusted means were similar, but the adjusted medians (from quantile regression) were approximately 2 g larger than the unadjusted medians.

For caffeine (with an intermediate degree of skewness in the dataset), the adjusted estimates obtained from linear regression and quantile regression were very different. In fact, the adjusted means (32.8 mg/day and 18.0 mg/day) from linear regression were shifted substantially to the right of the adjusted medians (15.3 mg/day and 8.4 mg/day) from quantile regression. In addition, linear regression gave significance at the 5% level, whereas quantile regression did not. The adjusted estimates from log-normal regression were substantially smaller than those from quantile regression, which again suggests that the log transformation might have skewed the distribution slightly to the left. The adjusted (from linear regression) and unadjusted means were similar, but the adjusted medians (from quantile regression) were approximately 2 mg/day larger than the unadjusted medians.

Finally, for vitamin K (largest skewness in the dataset), the adjusted estimates from linear regression and quantile regression were very different. The adjusted means (199.4 μg/day and 147.2 μg/day) from linear regression were again shifted substantially to the right of the adjusted medians (130.3 μg/day and 105.7 μg/day) from quantile regression. Furthermore, the linear regression approached significance ($P = 0.053$), whereas quantile regression did not. The adjusted estimates from log-normal regression were similar to those from quantile regression, suggesting that the log-normal regression might have transformed the distribution to normality. Adjustment again had a larger impact on the median; the unadjusted median for blacks (123.5 μg/day) was higher than for whites (117.6 μg/day), but the use of quantile regression changed those values substantially to 105.7 μg/day for blacks and 130.3 μg/day for whites.

This study illustrates the use of nonparametric median regression methods to obtain adjusted medians. We found that the use of median regression, linear regression, and log-normal regression yielded substantially different results; these differences highlight the discernibly different, and possibly conflicting, results that can be produced with parametric assumptions (e.g., normal or log-normal) compared to nonparametric median regression. Simulation studies have demonstrated that the use of the wrong transformation of the outcome can give very biased estimates of the median, whereas median regression gives unbiased results *(14)*. Because the correct transformation to achieve normality is unknown for any given outcome variable, nonparametric median regression can be used as an alternative.

Quantile regression can be used for continuous outcome data that may not be normally distributed, but this method should not be applied indiscriminately. For example, quantile regression is not appropriate for discrete ordinal data, for which proportional odds logistic regression may be appropriate; also, it may not be appropriate for multimodal data. If only a comparison of central tendency across groups is needed, then comparison of adjusted medians may be of primary interest. If the full distribution across groups must be compared, then comparison of adjusted 25th, 50th, and 75th percentiles may be of interest. In this case, conflicting results may be obtained. For example, the adjusted 25th percentiles could be similar across groups, but the adjusted medians could be different. The reporting should make it clear that the groups are similar in the lower quartile, but different in the center.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the Disclosures of Potential Conflict of Interest form. Potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: None declared.
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: The authors are grateful for the support provided by the following grants from the NIH: AI 60373, CA 74015, CA 069222, and MH 054693. J. A. Linder, American Medical Association, Roche Pharmaceuticals, and Pfizer.
Expert Testimony: None declared.

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, preparation or approval of manuscript.

References

1. Bassett G, Koenker R. An empirical quantile function for linear models with iid errors. J Am Stat Assoc 1982;77:407–15.
2. Friedrich N, Alte D, Völzke H, Spilcke-Liss E, Lüdemann J, Lerch MM, et al. Reference ranges of serum IGF-1 and IGFBP-3 levels in a general adult population: results of the Study of Health in Pomerania (SHIP). Growth Horm IGF Res 2008; 18:228–37.
3. Hewson P. Quantile regression provides a fuller analysis of speed data. Accid Anal Prev 2008;40: 502–10.
4. Lykoudis S, Psounis N, Mavrakis A, Christides A. Predicting photochemical pollution in an industrial area. Environ Monit Assess 2008;142:279–88.
5. Ovbiagele B, Buck BH, Liebeskind DS, Starkman S, Bang OY, Ali LK, et al. Prior antiplatelet use and infarct volume in ischemic stroke. J Neurol Sci 2008;264:140–4.
6. Rühli F, Henneberg M, Schaer DJ, Imhof A, Schleif-

fenbaum B, Woitek U. Determinants of inter-individual cholesterol level variation in an unbiased young male sample. Swiss Med Wkly 2008; 138:286–91.
7. TeMoananui R, Kieser JA, Herbison GP, Liversidge HM. Estimating age in Maori, Pacific Island, and European children from New Zealand. J Forensic Sci 2008;53:401–4.
8. Yi M, Meric-Bernstam F, Ross MI, Akins JS, Hwang RF, Lucci A, et al. How many sentinel lymph nodes are enough during sentinel lymph node dissection for breast cancer? Cancer 2008; 113:30–7.
9. McGreevy KM, Hoel B, Lipsitz SR, Hoel DG. Impact of nutrients on insulin-like growth factor-I, insulin-like growth factor binding protein-3 and their ratio in black and white males. Public Health Nutr 2006;10:97–105.
10. McGreevy KM, Hoel B, Lipsitz SR, Bissada NK, Hoel DG. Racial and anthropometric differences in insulin-like growth factor I (IGF-I) and insulin-like growth factor binding protein-3 (IGFBP-3) levels. Urology 2005;66:587–92.
11. Stanford JL, Stephenson RA, Coyle LM, Cerhan J, Correa R, Eley JW, et al. Prostate cancer trends 1973–1995. Bethesda: National Cancer Institute; 1999.
12. Mardia K. Measures of multivariate skewness and kurtosis with applications. Biometrika 1980;57: 519–30.
13. Willett WC. Nutritional epidemiology. 2nd ed. New York: Oxford University Press; 1998. 528 pp.
14. Fitzmaurice GM, Lipsitz SR and Parzen M. Approximate median regression via the box-Cox transformation. Am Stat 2007;61:233–8.

DOI: 10.1373/clinchem.2008.106260