Negative Consequences of Dichotomizing Continuous Predictor Variables

Article *in* Journal of Marketing Research · August 2003

DOI: 10.1509/jmkr.40.3.366.19237

2 authors:

Gary H Mcclelland
University of Colorado Boulder
92 PUBLICATIONS **6,547** CITATIONS

SEE PROFILE

Julie R. Irwin
University of Texas at Austin
43 PUBLICATIONS **2,737** CITATIONS

SEE PROFILE

JULIE R. IRWIN and GARY H. McCLELLAND*

Marketing researchers frequently split (dichotomize) continuous predictor variables into two groups, as with a median split, before performing data analysis. The practice is prevalent, but its effects are not well understood. In this article, the authors present historical results on the effects of dichotomization of normal predictor variables rederived in a regression context that may be more relevant to marketing researchers. The authors then present new results on the effect of dichotomizing continuous predictor variables with various nonnormal distributions and examine the effects of dichotomization on model specification and fit in multiple regression. The authors conclude that dichotomization has only negative consequences and should be avoided.

Negative Consequences of Dichotomizing Continuous Predictor Variables

Suppose a market researcher wanted to measure the effect of consumer experience, as measured by a continuous familiarity scale, on brand preference. The researcher might first split the familiarity score into two groups, familiar and unfamiliar, before performing the data analyses. In this article, we address some of the consequences of dichotomizing continuous predictor variables in this way.

Dichotomization of this sort is commonly practiced in marketing and throughout the social sciences. In three recent issues of *Journal of Marketing Research*, *Journal of Marketing*, and *Journal of Consumer Research*, 46% of the articles in which a predictor variable is measured continuously include at least one categorization of a continuous predictor variable.[1] Almost all (88%) these instances are dichotomizations. Median splits are the most popular method of dichotomization (66%), though authors also employ other methods, such as picking a cutoff point or dividing at the middle of a scale. Many types of predictor variables are commonly dichotomized, including personality variables (e.g., need for cognition: high versus low), cognitive variables (e.g., expertise: expert versus novice), demographic variables (e.g., age: young versus old), and survey research variables (e.g., survey response time: early versus late).

AN ILLUSTRATION OF THE EFFECTS OF DICHOTOMIZATION

Before considering formal results, we present an illustrative example of the basic issue. We randomly sampled the criterion (Y) and predictor (X) data in the left-hand graph in Figure 1 from a bivariate normal distribution for which the population squared correlation was .25. Y represents a variable such as purchase intention for a microwave oven (1–50 scale), and X represents a variable such as number of extra microwave cooking settings (0–33 scale). The sample squared correlation of .27 is close to the population correlation and would be significant at conventional levels (t[18] = 2.58; p = .02). From these data, it could be correctly concluded that the number of additional cooking settings significantly affects purchase intentions for microwave ovens.

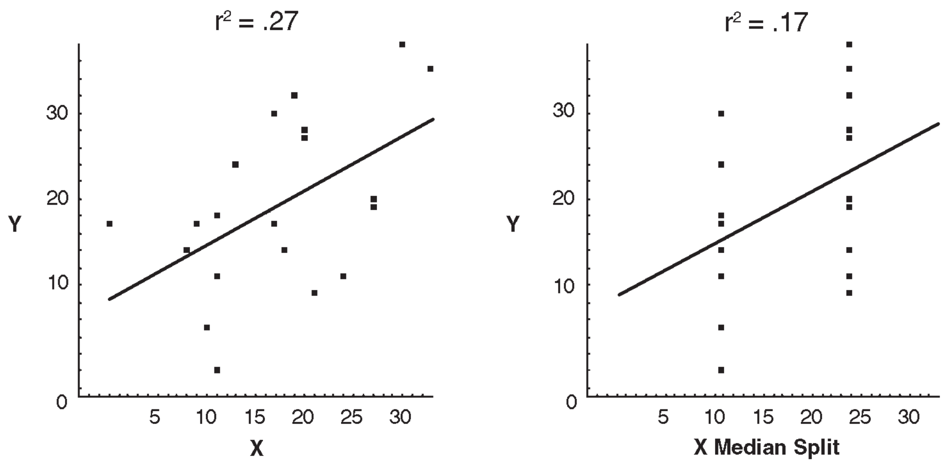
The right-hand graph in Figure 1 regresses the same Y variable on X dichotomized via a median split (median = 17.5). Usually, when researchers dichotomize their data in this way, they code the categorical data using dummy codes (0,1) before performing a regression, or they run an analysis of variance program that codes the data automatically. Because the choice of coding has no effect on the significance tests, and because it is easier to interpret the slopes if the scales stay constant across models, we used the mean value of X for the low and high groups, respectively, as the codes in the regression using the median-split data (on the right-hand side in Figure 1). We refer to this split variable (and other split variables in the sequel) as X′. The squared correlation between Y and X′, the split form of X, is .17, approximately 63% of the original squared correlation and no longer significant at conventional levels (t[18] = 1.89; p =

[1] See *Journal of Marketing Research*, 38 (November), 39 (February), and 39 (May); *Journal of Marketing*, 65 (October), 66 (January), and 66 (April); *Journal of Consumer Research*, 28 (2), 28 (3), and 28 (4).

*Julie R. Irwin is Associate Professor of Marketing, University of Texas at Austin (e-mail: Julie.Irwin@bus.utexas.edu). Gary H. McClelland is Professor of Psychology and Marketing, University of Colorado (e-mail: Gary.McClelland@colorado.edu). The authors thank the three anonymous *JMR* reviewers and Leigh McAlister for their helpful comments.

Figure 1

REGRESSIONS OF Y ON CONTINUOUS (LEFT) AND DICHOTOMIZED (RIGHT) X



.07). Dichotomization of the predictor variable X substantially reduces the squared correlation with Y from a significant value to a nonsignificant value. Regression of Y on X produces the slope estimate of .63 with a standard error of .24. Regression of Y on X′ produces essentially the same slope estimate of .61 but with a substantially larger standard error of .32. A researcher who obtains these results may conclude that the number of cooking settings does not have a significant effect on intentions to purchase microwave ovens.

REDERIVATION OF CLASSIC RESULTS

Previous treatments of this issue (e.g., Pearson 1900; Tate 1955) have not been absorbed by the marketing literature for three reasons. First, the earlier proofs did not take a modern regression approach. Second, the proofs are not particularly accessible either mathematically or in their exposition. Third, and most important, these previous articles provide solutions to a different problem from the median-split issue of interest today. The previous work focuses on estimating the exact correlation between Y and the latent continuous variable underlying an X variable measured discretely. Modern researchers more likely focus on the reverse problem of how much the correlation with a continuous predictor is deflated if a dichotomization of that predictor is used instead. We analyze the deflation in the context of correlation and regression to show how familiar components of regression results change when a continuous predictor is dichotomized. An important benefit of deriving the results in a regression context is that it is easy to generalize the results for the first time to dichotomizing various nonnormal distributions for predictor variables that are likely to arise in research.

Assume there is a linear relationship between predictor variable X and criterion variable Y:

$$Y_i = a + bX_i + e_i. \tag{1}$$

According to Cohen and Cohen (1983, p. 42), the squared correlation between X and Y is given by

$$r^2_{XY} = \frac{b^2\sigma^2_X}{\sigma^2_Y}. \tag{2}$$

To consider the effects that dichotomizing X at its median would have on the squared correlation between the dichotomized variable and Y, we must first consider the effects of dichotomizing on the distribution of X itself. If f(X) represents the probability density function over X, then the expected or mean values of X for below and above the median are given by

$$\mu_L = 2\int_{-\infty}^{q_{.5}} Xf(X)dX, \text{ and } \mu_H = 2\int_{q_{.5}}^{\infty} Xf(X)dX, \tag{3}$$

where $q_{.5}$ represents the .5 quantile or median. For example, consider the standard normal distribution with mean 0 and standard deviation 1. The corresponding mean values of X for the two groups, one greater than and one less than the median of 0, are given by Equation 3 after substituting the formula for the probability density function of the standard normal distribution:

$$\mu_L = \int_{-\infty}^{0} 2X\frac{1}{\sqrt{2\pi}}e^{-X^2}dX = -\sqrt{\frac{2}{\pi}} \approx -.798, \text{ and} \tag{4}$$

$$\mu_H = \int_{0}^{\infty} 2X\frac{1}{\sqrt{2\pi}}e^{-X^2}dX = \sqrt{\frac{2}{\pi}} \approx +.798.$$

Dichotomization of X at its median is analogous to creating a density function g(X′) with all the probability mass at two points, the means of the upper and lower groups. The variance of X′ for the dichotomized density function g(X′) is

$$\sigma^2_{X'} = \frac{(\mu_H - \mu_L)^2}{4}. \tag{5}$$

For linear relationships as defined by Equation 1, the regression line passes through the point determined by the means of the two variables, so that the mean of the criterion variable equals the intercept plus the slope times the mean of the predictor variable (Judd and McClelland 1989, p. 115). Indeed, for any subset of the data, the nature of a linear relationship ensures that

$$\bar{Y}_i = a + b\bar{X}_i + \bar{e}_i. \tag{6}$$

Therefore, if we instead use a dichotomous predictor variable X′ that has the value of the mean of the X for those observations above the median and has the value of the mean of those observations for those observations below the median, then

$$Y_i = a + bX'_i + k\varepsilon_i, \tag{7}$$

where k is a factor (greater than 1) representing the increase in error due to the aggregation of observations with disparate values of the predictor into two groups (in the previous data example, k = 1.33). Thus, if the within-group predictor mean is the dichotomous score (as opposed to dummy or some other codes), regression of the dependent variable on either the continuous or the dichotomized predictor will produce the same estimate of the regression slope and intercept. On average, the mean squared error will be larger for the dichotomous analysis than for the continuous analysis with a concomitant loss of power and wider confidence interval.

Given that the slope remains unbiased with dichotomization, we can substitute the variance of the dichotomized predictor (Equation 7) in the formula for the squared correlation (Equation 2) to determine the expected value after dichotomization. That is, the squared correlation using the split variable is given by

$$r^2_{\text{split}} = \frac{b^2\sigma^2_{X'}}{\sigma^2_Y} = \frac{b^2(\mu_H - \mu_L)^2}{4\sigma^2_Y}. \tag{8}$$

Thus, the ratio of the squared correlations for the dichotomized and original values of the predictor X equals the ratio of Equations 8 and 2. The ratio of the squared correlations is determined by two factors: (1) the mean value of the dependent variable in the original data for the two dichotomized groups and (2) the original variance of X. Note that the ratio does not depend on the distribution of X. Nevertheless, it is worthwhile to compute the ratio for a predictor variable with a standard normal distribution. Substitution of the two group means resulting from dichotomization of a standard normal predictor X (Equation 4) into the ratio of the squared correlations for the dichotomized and original values of X yields

$$\text{ratio} = \frac{r^2_{\text{split}}}{r^2} = \frac{(\mu_H - \mu_L)^2}{4\sigma^2_X} = \frac{\left[\sqrt{2/\pi} - \left(-\sqrt{2/\pi}\right)\right]^2}{4} = \frac{2}{\pi} \approx .637. \tag{9}$$

The squared correlation becomes 64% of what it would have been. This is the same numerical value that Peters and Van Voorhis (1940), Tate (1955), and Srinivasan and Basu (1989) obtain in different contexts. Our derivation of this result in a regression context enables us to calculate the reduction in the squared correlation between X and Y when the normally distributed X is split at different points other than the median. We calculated the reduction at the 60th, 70th, 80th, 90th, and 95th percentiles, and the resulting squared correlations were 62%, 58%, 49%, 32%, and 22%, respectively, of the original squared correlations. These results show that the median is the most powerful place to split the data, but even then the power loss is quite costly.

DICHOTOMIZATION OF VARIABLES WITH SKEWED AND BIMODAL DISTRIBUTIONS

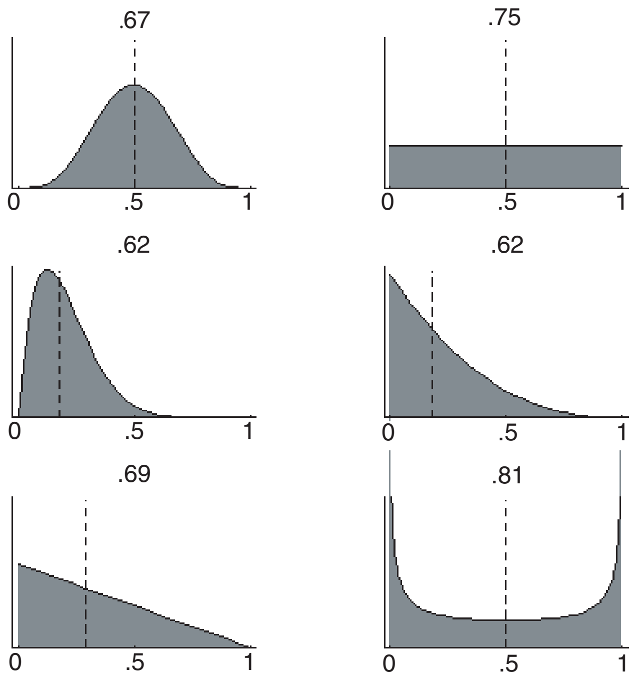
A more worthwhile application of our regression derivations is to help better understand the effects of dichotomizing nonnormal distributions, because many predictor variables in marketing are likely to have a nonnormal distribution. In Figure 2, we consider the effects of splitting various beta distributions at their medians. We use beta distributions because they can produce several distributional shapes, and in contrast with normal distributions, they have the more realistic assumption of finite range. Consider the beta distribution whose probability density is given by

$$f(X) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} X^{p-1}(1-X)^{q-1}, \quad 0 < X < 1. \tag{10}$$

The distribution of X is symmetric when p = q and is skewed otherwise. When p = q = 1, the beta distribution is equivalent to the uniform distribution for which every value between 0 and 1 is equally likely. Values of p and q greater than 1 result in inverted U-shaped or unimodal distributions,

Figure 2

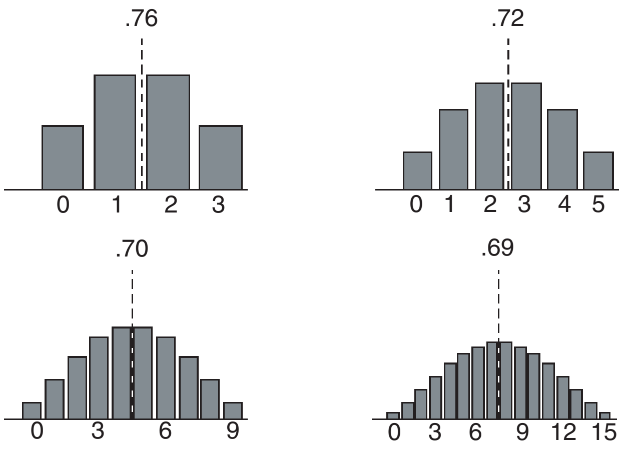
VARIOUS SHAPES OF THE BETA DISTRIBUTION AND THE CORRESPONDING EXPECTED REDUCTION OF THE SQUARED CORRELATION BY SPLITTING AT THE MEDIAN



Notes: Parameter values p and q, from left to right and top to bottom, are (5,5), (1,1), (2,8), (1,3.4), (1,2), and (.5,.5).

Figure 3

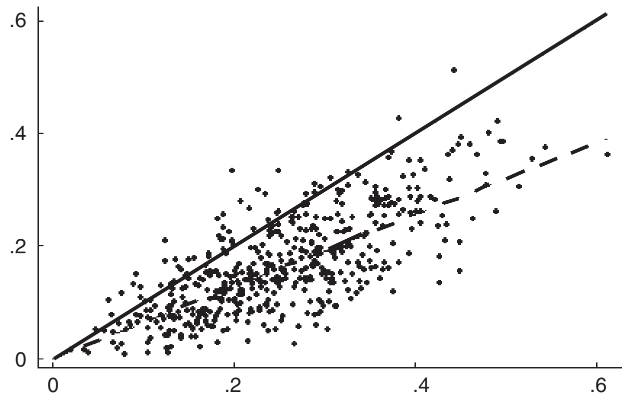
VARIOUS MULTISTEP DISTRIBUTIONS AND THE CORRESPONDING EXPECTED REDUCTION OF THE SQUARED CORRELATION BY SPLITTING AT THE MEDIAN



Notes: Beta-binomial distributions with parameters (3,3).

Figure 4

SAMPLING DISTRIBUTION FOR THE SQUARED CORRELATION COEFFICIENT FOR A BIVARIATE NORMAL DISTRIBUTION



Notes: The population squared correlation coefficient is .25; each point represents a sample size of 50. The x-axis is the squared correlation calculated by means of a continuous predictor variable, and the y-axis is the squared correlation calculated by means of a median split of the predictor variable.

whereas values of p and q less than 1 result in U-shaped or bimodal distributions. Figure 2 depicts various beta distributions and the corresponding proportional reduction in the squared correlation due to dichotomization of the predictor at the median. We obtained these results by substituting Equation 10 for f(X) in Equation 4.

Dichotomization of the symmetric beta distribution with shape parameters p = 5 and q = 5 (a symmetric, normal-like distribution) reduces the expected squared correlation to 67% of what it would have been, about the same as the approximate 64% for the median split of the normal distribution. Note that in Figure 2, the only distributions that fare better than the normal distribution after dichotomization are distributions with substantial probabilities in both tails, the uniform and the bimodal, with expected reductions to 75% and 81%, respectively, of the original squared correlations. Dichotomization of variables by splitting them at the median never produces an improvement in the expected squared correlation between X and Y for the various distributional shapes in Figure 2.

DICHOTOMIZATION OF MULTISTEP VARIABLES WITH SYMMETRIC DISTRIBUTIONS

Researchers seldom use true continuous variables; instead, they use multistep variables, such as steps on a rating scale. Figure 3 shows these effects for normal-like, symmetric, multistep distributions that vary in the number of steps. The fewer steps there are, the less is the proportional reduction in the squared correlation due to dichotomization. However, dichotomization of a symmetric distribution with only four steps reduces the expected squared correlation to approximately 76% of what it otherwise would have been, which suggests that dichotomization of symmetric, multistep variables is only slightly less damaging than dichotomization of normally distributed variables. This result is consistent with Srinivasan and Basu's (1989) conclusion that ratings scales with enough steps approximate the metric quality of true continuous scales. Figure 3 provides the additional insight that a median split is not appropriate even when the original predictor variable scale only has four steps.

For brevity, we do not present graphs for multistep variables with asymmetric or bimodal distributions. The results are comparable to the results for continuous data. For an eight-step variable, the percentages for dichotomization of a beta-binomial variable with the same beta shape parameters as those in Figure 2 are (left to right and top to bottom) 69%, 76%, 58%, 60%, 67%, and 82%.

EXPECTED VERSUS ACTUAL MODEL FIT

All our previous and subsequent demonstrations necessarily involve expected values for the squared correlation. Figure 4 shows an empirical sampling distribution for the squared correlation coefficient for a bivariate normal distribution for which the population squared correlation coefficient (i.e., the actual relationship between the two variables) is .25. Each point represents a sample size of 50. The horizontal axis is the squared correlation calculated using the continuous predictor variable, and the vertical axis is the squared correlation calculated using a median split of the predictor variable. Thus, points below the 45-degree diagonal line represent samples in which splitting decreases the squared correlation, and points above the line represent samples in which splitting increases the squared correlation. The broken diagonal line represents the expected 64% reduction in the squared correlation as a result of splitting one of the two variables. The points are scattered about that line; the degree of scatter depends on the sample size. Because the line represents an expected value, sometimes dichotomization reduces the squared correlation even more than expected, though at best it only slightly increases the squared correlation in a few scattered instances. The few points that do show an increase in squared correlation may seem like a boon to researchers: The dichotomization apparently increases power. However, obtaining a significant result with a dichotomous predictor variable when the original continuous variable is not significant does not mean that the model using the median split is more accurate, but rather that, because of sampling error, the analysis has resulted in an inaccurately high estimate of the true correlation.

DICHOTOMIZATION AND STATISTICAL SIGNIFICANCE

Related to the expected decrease in the squared correlation after dichotomization of a predictor variable is the effect of dichotomization on statistical significance of the ensuing tests. One motivation for dichotomizing a continuous variable is to form two groups in order to run an analysis of variance or to perform a t-test. The formula (e.g., Rosenthal and Rosnow 1991, p. 317) for relating the squared correlation (r^2) to the corresponding F is

$$F_{1,\mathrm{d.f.}} = \frac{r^2}{1 - r^2}\,\mathrm{d.f.} \tag{11}$$

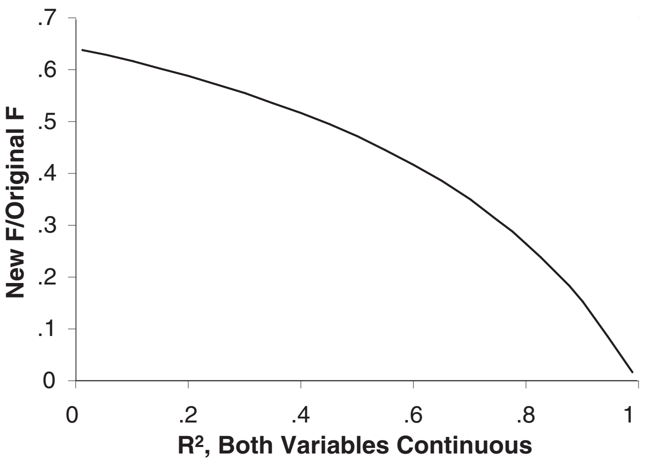
The expected value of F when using a dichotomized variable can be obtained by substituting the reduced value for the squared correlation for the appropriate distribution from Equation 11. For example, dichotomization of a normal distribution reduces the squared correlation by approximately 64%; thus, the corresponding F would be

$$F_{1,\mathrm{d.f.}} = \frac{.64r^2}{1 - .64r^2}\,\mathrm{d.f.} \tag{12}$$

Figure 5 displays this expected reduction in the F statistic as a function of the squared correlation for the continuous variable. The stronger the relationship among the original variables, the greater the deleterious effect is of dichotomizing on the F statistic. For example, for an original squared correlation of .5, dichotomization reduces F to about 47% of what it would have been; for an original squared correlation of .75, the reduction is to about 31% of the original F. Whether the decreased F was still significant at a given alpha level would depend on sample size, but the overall dichotomization substantially reduces the chances of statistical significance.

Figure 5

EXPECTED REDUCTION IN F DUE TO MEDIAN SPLIT AS A FUNCTION OF THE ORIGINAL SQUARED CORRELATION BETWEEN THE CONTINUOUS VARIABLES



DICHOTOMIZATION IN MULTIPLE REGRESSION

In simple regression, dichotomization reduces power to detect relationships that are in the data. In this case, dichotomization does not bias the estimation of the model parameters (the slope and the intercept); rather, it adds error to their estimation. In contrast, in multiple regression, dichotomization of several continuous predictor variables can exacerbate or disguise model misspecification problems and can induce models that do not accurately reflect the underlying data.

Most researchers who perform multiple regression are interested in the model results with all the predictors in the model (e.g., whether one predictor variable is significant when another predictor variable is included in the model). For example, causal claims are made more easily when mediation or suppression (Baron and Kenny 1986) has been observed. In a notable stream of research, Maxwell and Delaney (1993) and Vargha and colleagues (1996) show that after dichotomization, the multivariate relationships among predictor variables actually can be obscured. Predictor variables can appear significant when they are not significant in the original data.

The intuition behind the findings is consonant with our derivations thus far (the actual derivations can be found in the original articles). Dichotomization suppresses simple relationships among variables, and the relationships between correlated predictor variables are similarly affected. Sometimes the reduction in the simple correlations among the predictors and the criterion variable, and among one another, can result in a misleading multiple regression model. If the collinearity is artificially suppressed, one predictor variable may appear significant in the multiple regression when it is dichotomized but not when it is not dichotomized (i.e., a spurious effect).

CONSERVATISM AND TYPE II ERROR

Of the many effects of dichotomization, power reduction seems the most innocuous and could be presented as a benefit, because it renders the analysis "conservative." For scientists who wish to avoid misleading anyone with inaccurately significant results, Type II errors (i.e., not finding a significant result when the effect is present) seem less costly than Type I errors (i.e., finding a significant result when the effect is not present).

The use of "conservatism" in this sense is problematic on three counts: (1) The addition of error to data can sometimes result in spurious effects as well as a dampening of the actual effects; (2) when a significant effect is found using a weak analysis technique, that effect represents a waste of resources, because fewer observations could have been used to obtain the same result; and (3) when weak analysis tools are used, actual effects are likely to remain undiscovered because of resource constraints. This last problem is especially notable when the appropriate data are field data (a common situation in marketing) and/or data using rarer subject groups. The marketing community arguably has less information about important market segments (e.g., children, couples, lower-income workers, minorities) partly because it is so difficult to find and use these groups as subjects. The reduction of statistical power to view true relationships in these data is undesirable.

CONCLUSION

Dichotomization of predictor variables substantially reduces power in simple regression. This loss of power applies not only to normal distributions but also to skewed and bimodal continuous distributions and to multistep distributions (even those with as few as four steps). The effect on the squared correlation between the predictor and criterion variables as well as on the significance levels of the predictor variable is substantial and, on average, affects the ability to detect actual relationships. For multiple regression, the situation is more complicated, but it is still negative. Dichotomization can distort the true relationships among collinear predictor variables and can impede the selection of the appropriate multivariate model.

In conclusion, dichotomization of predictor variables has serious costs and no benefits. It is prevalent in marketing (and in other applied and social sciences). It is an undesirable practice that should no longer be used in research.

REFERENCES

Baron, R.M. and David A. Kenny (1986), “The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations,” *Journal of Personality and Social Psychology*, 51 (6), 1173–82.

Cohen, Jacob and Patricia Cohen (1983), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Judd, Charles M. and Gary H. McClelland (1989), *Data Analysis: A Model-Comparison Approach.* San Diego: Harcourt Brace Jovanovich.

Maxwell, Scott E. and Harold D. Delaney (1993), “Bivariate Median Splits and Spurious Statistical Significance,” *Psychological Bulletin,* 113 (1), 181–90.

Pearson, Karl (1900), *Mathematical Contributions to the Theory of Evolution, VII: On the Correlation of Characters Not Quantitatively Measurable*, Philosophical Transactions of the Royal Society of London, Series 195A, 1–47.

Peters, Charles C. and Walter R. Van Voorhis (1940), *Statistical Procedures and Their Mathematical Bases.* New York: McGraw-Hill.

Rosenthal, Robert and Ralph L. Rosnow (1991), *Essentials of Behavioral Research: Methods and Data Analysis*, 2d ed. New York: McGraw-Hill.

Srinivasan, V. and Ayima K. Basu (1989), “The Metric Quality of Ordered Categorical Data,” *Marketing Science*, 8 (3), 205–230.

Tate, Robert F. (1955), “The Theory of Correlation Between Two Continuous Variables When One Is Dichotomized,” *Biometrika*, 42 (2), 205–216.

Vargha, Andras, Tamas Rudas, Harold D. Delaney, and Scott E. Maxwell (1996), “Dichotomization, Partial Correlation, and Conditional Independence,” *Journal of Educational and Behavioral Statistics*, 21 (2), 264–82.