Available online at www.sciencedirect.com

ScienceDirect

Journal of Consumer Psychology 25, 4 (2015) 666 – 678

JCP JOURNAL OF CONSUMER PSYCHOLOGY

Research Dialogue

A researcher's guide to regression, discretization, and median splits of continuous variables☆

Derek D. Rucker [a,*], Blakeley B. McShane [a], Kristopher J. Preacher [b]

[a] *Marketing Department, Kellogg School of Management, Northwestern University, USA*
[b] *Department of Psychology & Human Development, Vanderbilt University, USA*

Received 24 April 2015; accepted 27 April 2015
Available online 6 May 2015

Abstract

We comment on Iacobucci, Posavac, Kardes, Schneider, and Popovich (2015) by evaluating the practice of discretizing continuous variables. We show that dichotomizing a continuous variable via the median split procedure or otherwise and analyzing the resulting data via ANOVA involves a large number of costs that can be avoided by preserving the continuous nature of the variable and analyzing the data via linear regression. As a consequence, we recommend that regression remain the normative procedure both when the statistical assumptions explored by Iacobucci et al. hold and more generally in research involving continuous variables. We also discuss the advantages of preserving the continuous nature of the variable for graphical presentation and provide practical suggestions for such presentations.

Keywords: Continuous; Discretization; Categorization; Dichotomization; Median split; Regression

Introduction

Iacobucci, Posavac, Kardes, Schneider, and Popovich (IPKSP) have brought the general issue of the discretization (or categorization) of continuous variables—and the specific issue of discretization via the median split procedure—to the forefront of consumer psychology. We appreciate their call for "a more nuanced understanding of the statistical properties of a median split." It is useful and constructive to discuss best practices in research, and we thank the editors of the *Journal of Consumer Psychology* for allowing us to offer our own perspective and contribution.

In this commentary, we provide a "researcher's guide" to regression, discretization, and median splits of continuous variables. Our commentary is targeted at both those who are unfamiliar with the core issues involved in discretization as well as those who wish to better understand them. To preview the perspective proffered in this commentary, it is helpful to recall the following statement from IPKSP:

> Although median splits may be perceived as suboptimal from the perspective of power, if there was no possibility that they could produce misleading support for apparent relations between variables that, in truth, are spurious, their use would not be a problem. However, if median splits can produce Type I errors (the false conclusion of an effect), their use would be inappropriate.

In contrast to IPKSP's near exclusive focus on Type I error, we propose a more holistic and integrative view of the costs and perceived benefits associated with discretization. In particular, we believe that Type II error is of considerable importance and suggest that the relative cost of Type I versus Type II error (which varies by research setting) should be given due consideration. Consequently, we emphasize the use of efficient statistical

☆ The authors thank Ulf Böckenholt for helpful comments on this manuscript.
* Corresponding author at: Marketing Department, Kellogg School of Management, Northwestern University, 2001 Sheridan Road, Evanston, IL 60208, USA.
E-mail address: d-rucker@kellogg.northwestern.edu (D.D. Rucker).

http://dx.doi.org/10.1016/j.jcps.2015.04.004
1057-7408

procedures in order to increase statistical power (i.e., decrease Type II error) while keeping Type I error fixed at α, the size of the test (i.e., the maximum probability of a Type I error or the minimum significance level; typically $\alpha = 0.05$). We also discuss additional costs of discretization such as the loss of individual-level variation, reduced predictive performance, and inefficient effect size estimates.

In the remainder of this commentary, we provide an in depth treatment of issues pertaining to the discretization of a continuous variable under the statistical assumptions explored by IPKSP. We explore the issues associated with various statistical approaches to continuous variables and discretization and show that preserving the continuous nature of the variable and analyzing the data via linear regression both is more informative and has greater power. As noted, the rationale for this perspective is based on more general considerations beyond the Type I error issues explored by IPKSP. We aim to make it transparent that discretization is associated with a large number of costs. Thus, we recommend that regression remain the normative procedure both when the statistical assumptions explored by IPKSP hold and more generally in research involving continuous variables. We also discuss the advantages of preserving the continuous nature of the variable for graphical presentation. Finally, we comment briefly on several additional considerations and provide a brief summation.

Analysis strategies in the default case

Statistical assumptions

IPKSP rightly note that it is quite common for a "researcher to manipulate one (or more) factors and measure another [continuous] variable (one but not more)" and thus we make this case the focus of our commentary. We further assume that the sample size per (manipulated) condition is not unreasonably small because (i) one generally does not attempt to account for a measured variable (whether by regression or other means) when one has very few subjects per condition and (ii) reasonable sample sizes ensure the measured variable is independent of the (generally randomized) treatment manipulation provided either the measured variable is assessed prior to the manipulation or it is assessed after the manipulation but is unaffected by it; this assumption does not seem unreasonable as the field is moving in the direction of larger sample sizes (Asendorpf et al., 2013; Brandt et al., 2014; Cumming, 2014; McShane & Böckenholt, 2014, in press; Pashler & Wagenmakers, 2012).

We further assume that the linear model holds. That is, for each manipulated condition, the continuous dependent variable Y is a linear function of the measured variable X.

The assumptions stated thus far are entirely consistent with those of IPKSP. We make the additional assumption that the number of manipulated conditions is two and they are labeled "Treatment" and "Control." As such, the treatment status can be represented by a binary treatment variable T that is zero for subjects in the control condition and one for those in the treated condition. We make this last assumption for illustration purposes only and our comments hold for any natural number of conditions (e.g., four in a 2 × 2 study design).

These assumptions define the "default case" for this commentary and we explicitly note whenever we make comments that depart from this case.

Linear regression

We illustrate linear regression in the default case in Fig. 1. The data underlying the figure comes from a simulated two-condition study with two hundred subjects per condition. The points indicate the raw data, the x-axis indicates the measured variable, the y-axis indicates the dependent variable, and the color indicates the treatment variable. Finally, the lines indicate a linear regression fit to the data; this can be fit by regressing the dependent variable Y on the treatment variable T, the measured variable X, and their product $T \cdot X$.

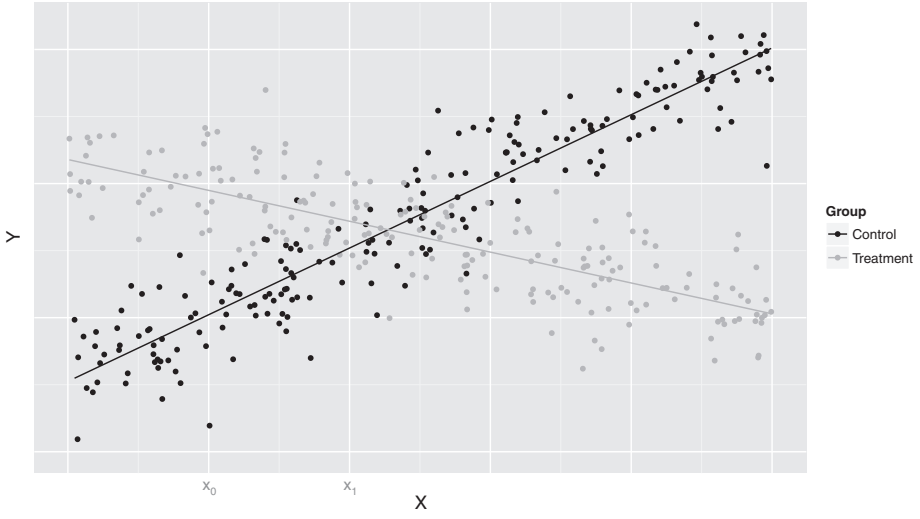

Fig. 1. Linear regression in the default case. The points indicate the raw data, the x-axis indicates the measured variable, the y-axis indicates the dependent variable, and the color indicates the treatment variable. The lines indicate a linear regression fit to the data.