Thoresen *BMC Medical Research Methodology* (2019) 19:28
https://doi.org/10.1186/s12874-019-0667-2

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Spurious interaction as a result of categorization

Magne Thoresen

Abstract

Background: It is common in applied epidemiological and clinical research to convert continuous variables into categorical variables by grouping values into categories. Such categorized variables are then often used as exposure variables in some regression model. There are numerous statistical arguments why this practice should be avoided, and in this paper we present yet another such argument.

Methods: We show that categorization may lead to spurious interaction in multiple regression models. We give precise analytical expressions for when this may happen in the linear regression model with normally distributed exposure variables, and we show by simulations that the analytical results are valid also for other distributions. Further, we give an interpretation of the results in terms of a measurement error problem.

Results: We show that, in the case of a linear model with two normally distributed exposure variables, both categorized at the same cut point, a spurious interaction will be induced unless the two variables are categorized at the median or they are uncorrelated. In simulations with exposure variables following other distributions, we confirm this general effect of categorization, but we also show that the effect of the choice of cut point varies over different distributions.

Conclusion: Categorization of continuous exposure variables leads to a number of problems, among them spurious interaction effects. Hence, this practice should be avoided and other methods should be considered.

Keywords: Categorization, Dichotomization, Interaction, Regression, Measurement error

Background

It is common in epidemiological and medical research to categorize exposure variables measured on a continuous scale and treat them as categorical in the statistical analysis. The continuous variables can be dichotomized or they can be divided into more than two groups, the latter alternative allowing investigation of a possible dose-response relationship. Examples of this practice include Body Mass Index (BMI) categorized according to pre-defined values and nutritional intake categorized according to observed quintiles. There may be several reasons for this practice. The most common ones being that it makes the analysis and the interpretation easier; one avoids having to model the actual relationship between the exposure variables and the response, and that it mimics clinical practice where one typically divides patients into groups (hypertensive vs. normotensive, obese vs. non-obese).

A number of papers have appeared, both in the biostatistical [1–7], epidemiological [8–13] and psychological [14–16] literature, pointing to problems with this approach and arguing against it. Among the problems are loss of information and power, but also an increased risk of type I error if continuous confounder variables are categorized. Recently, also predictive performance of models with categorized predictors is criticized [7]. We will not repeat their arguments here, but rather point to a problem that has received much less attention; that categorization of continuous exposure variables may lead to spurious interaction effects in a multiple regression model against an outcome.

This problem was observed already in 1974, by Humphreys and Fleishman [17], in a simulation study of the behavior of the ANOVA model in situations with two categorized explanatory variables. Later, the same

Correspondence: magne.thoresen@medisin.uio.no
Centre for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, P.O. Box 1122, Blindern, N-0317 Oslo, Norway

type of problem was also noted by Paunonen and Jackson [18] in an investigation of possible effect modification of the association between personality trait measures and resulting behavior. They noted that effect modification, or interaction, was often observed if the effect modifier in question was dichotomized and a stratified analysis was performed, while if the same association was investigated in a linear regression model, keeping the effect modifier on its original scale and introducing a product term, effect modification was less often observed. This observation led Bissonnette et al. [19] to perform a simulation study investigating the same problem. They also found that when they simulated a model with no interaction, dichotomized the potential effect modifier and then performed a stratified analysis, effect modification was often observed. However, no real understanding of or explanation for the findings was provided.

Maxwell and Delaney [14] carried out a formal analytical investigation of the effects of dichotomizing the explanatory variables in a linear regression model. Specifically, they showed that in a model with two correlated explanatory variables X_1 and X_2, if the true relationship between one of the explanatory variables and the outcome Y was non-linear, a spurious interaction effect appeared when dichotomizing X_1 and X_2 at the median. This result has later been referred to by a number of other authors who have discussed the practice of categorization [2, 16], and the non-linear nature of the relationship between the explanatory variable and the outcome seems to have been taken as the explanation of the spurious interaction.

In this paper we will look into this problem in some more detail. We have two exposure variables X_1 and X_2 that are correlated, and where both of them are dichotomized. The situation where only one of them is dichotomized follows directly. Furthermore, they are both to be related to an outcome variable Y by some regression model. Throughout we will assume that there is no interaction between X_1 and X_2. In much earlier work, it has been assumed that the variables have been categorized at the median. However, in many medical and epidemiological applications, it is more relevant to consider categorization at more extreme values, and / or at several cut points. This leads to some interesting findings.

In our analytical treatment of the problem, we take the regression model to be linear, and we assume normally distributed variables X_1 and X_2. However, the results apply to regression models and distributions in general. We have also carried out a small simulation study in order to explore the effects for different distributions. We will first give two examples to show the relevance of the problem.

Illustration 1, height and lung function

The first illustration uses data on lung function collected among Norwegian medical students. Peak expiratory flow (PEF), l/min, was measured six times for each student; three times in sitting position and three times in standing position. We will be using the mean of the six measurements in this illustration. In addition to PEF, we also measured the height of the students (cm) and we have gender information. In total we have data from 377 students, and we will model PEF as a function of gender and height (centered). Running the simple linear regression model $\text{PEF} = \beta_0 + \beta_1 \times \text{gender} + \beta_2 \times \text{height} + \beta 3 \times \text{gender} \times \text{height} + \varepsilon$ leads to the following estimated coefficients (SE): $\hat{\beta}_0$= 556.7 l/min (9.2 l/min), $\hat{\beta}_1$= – 129.7 l/min (11.0 l/min), $\hat{\beta}_2$ = 3.7 (l/min)/cm (0.9 (l/min)/cm), $\hat{\beta}_3$ = – 0.8 l/min (1.1 l/min). Using the conventional 5% significance level, we have clearly significant effects of gender and height, but no interaction. Next, we categorize height according to the gender specific 90th percentiles; 189 cm for men and 175 cm for women, coding zero if the subject is below the cut-off and one if above. We will then run the same linear model as above, but with the categorized version of height. This leads to the following estimated coefficients (SE): $\hat{\beta}_0$=578.4 l/min (6.4 l/min), $\hat{\beta}_1$ = – 168.8 l/min (8.1 l/min), $\hat{\beta}_2$ = 64.7 l/min (17.3 l/min), $\hat{\beta}_3$ = – 48.5 l/min (22.6 l/min). We notice that we have a significant interaction at the conventional 5% level ($p = 0.03$). This would indicate that the effect of being among the higher 10% is significantly lower among women than among men, and while being among the 10% highest males leads to an increased PEF of 64.7 l/min, the same increase is only 16.2 l/min for females.

Illustration 2, myocardial infarction

The second example is taken from a huge Norwegian health survey. During the period from 1985 until 1999 the Norwegian government conducted health surveys inviting men and women in the age of 40–42 years to participate. We will be analyzing a subset of these data, collected during the period 1985 to 1994. We have measured, among other things, Body Mass Index (BMI) and systolic blood pressure (BP) on a total of 133,139 subjects. These subjects have been followed for on average 19 years, and death of myocardial infarction is registered through a linkage to the Norwegian Cause of Death Registry. We will restrict our analysis to subjects with BMI > 20 to avoid having to deal with obvious non-linearities. This leaves us with 132,150 subjects. Among these there were 2542 (1.9%) deaths. We fit a logistic regression model for the odds of death by myocardial infarction as a linear function of BMI (kg/m^2) and BP (per 10 mmHg) and the interaction between BMI and BP. The estimated coefficients (SE) are 0.20 (0.04) for BMI, 0.61 (0.08) for BP, – 0.01 (0.003) for the interaction term and a constant term of – 14.06 (1.20). Due to the large sample size, all the

estimated effects are clearly significant at the conventional 5% level. However, the estimated interaction term has no practical significance. Next, we divide the sample into two BMI groups (obese vs. non-obese) by making a cut-off at 30 kg/m^2, and we divide the sample into two BP groups (hypertensive vs. normotensive) by making a cut-off at 140 mmHg. Based on the two categorized variables we run the same logistic model as above (main effects of BMI and BP and the interaction BMI × BP). The estimated coefficients are 0.81 (0.09) for BMI, 1.03 (0.04) for BP, – 0.41 (0.11) for the interaction and a constant term of – 4.43 (0.03). This would indicate that the effect of being obese varies between normotensive and hypertensive so that among normotensive subjects there is an odds ratio of death equal Exp(0.81) = 2.25, while among the hypertensive the effect of being obese is reduced to an odds ratio of Exp(0.81–0.41) = 1.49, a substantial difference.

Both examples show how categorization may lead to substantial changes in the interpretation of the data in practical data analysis.

Methods

Analytical developments, categorization in the bivariate normal situation

Assume we have a linear relationship between two exposure variables X_1, X_2, and an outcome variable Y, satisfying the linear regression equation.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \tag{1}$$

where we assume $(X_1, X_2) \sim N(0, \Sigma)$, $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$. Notice that there is no interaction in the true model. Furthermore, assume $\tilde{X}_1 = I(X_1 > c_1)$, $\tilde{X}_2 = I(X_2 > c_2)$ where for simplicity we let $c_1 = c_2 = c$. Let us define $\mu_{ij} = E(Y|\tilde{X}_1 = i, \tilde{X}_2 = j)$, $i, j = 0, 1$. We have no interaction between $\tilde{X}_1$ and $\tilde{X}_2$ if and only if $\mu_{00} - \mu_{10} = \mu_{01} - \mu_{11} \Rightarrow \mu_{11} - \mu_{01} - \mu_{10} + \mu_{00} = 0$.

Based on model (1) we have

$$\begin{aligned}\mu_{00} &= \beta_0 + \beta_1 E(X_1|X_1 \le c, X_2 \le c) \\ &\quad + \beta_2 E(X_2|X_1 \le c, X_2 \le c)\end{aligned} \tag{2}$$

$$\begin{aligned}\mu_{01} &= \beta_0 + \beta_1 E(X_1|X_1 \le c, X_2 > c) \\ &\quad + \beta_2 E(X_2|X_1 \le c, X_2 > c)\end{aligned} \tag{3}$$

$$\begin{aligned}\mu_{10} &= \beta_0 + \beta_1 E(X_1|X_1 > c, X_2 \le c) \\ &\quad + \beta_2 E(X_2|X_1 > c, X_2 \le c)\end{aligned} \tag{4}$$

$$\begin{aligned}\mu_{11} &= \beta_0 + \beta_1 E(X_1|X_1 > c, X_2 > c) \\ &\quad + \beta_2 E(X_2|X_1 > c, X_2 > c).\end{aligned} \tag{5}$$

In order to further investigate the relationships of interest, we need to be able to calculate the conditional expectations that enter these expressions. Define $F_{00} = P(X_1 \le c \cap X_2 \le c)$, $F_{01} = P(X_1 \le c \cap X_2 > c)$, $F_{10} = P(X_1 > c \cap X_2 \le c)$, $F_{11} = P(X_1 > c \cap X_2 > c)$, the probabilities of belonging to each of the four combinations of $\tilde{X}_1$, $\tilde{X}_2$. Due to symmetry,

$$\begin{aligned}E(X_1|X_1 > c, X_2 > c) &= E(X_2|X_1 > c, X_2 > c), \\ E(X_1|X_1 \le c, X_2 > c) &= E(X_2|X_1 > c, X_2 \le c), \\ E(X_1|X_1 > c, X_2 \le c) &= E(X_2|X_1 \le c, X_2 > c), \\ E(X_1|X_1 \le c, X_2 \le c) &= E(X_2|X_1 \le c, X_2 \le c),\end{aligned}$$

so in the following we will focus on the conditional expectations of X_1. Regier and Hamdan [20], using the Mehler identity [21], gave the following identity:

$$F_{11}E(X_1|X_1 > c, X_2 > c) = \phi(c)\left[1 - \Phi\left(\frac{c - \rho c}{\sqrt{1-\rho^2}}\right)\right](1 + \rho), \tag{6}$$

where $\phi(\cdot)$ denotes the normal density function and $\Phi(\cdot)$ denotes the corresponding cumulative distribution function.

As indicated by Regier and Hamdan [20], corresponding identities can be found by appropriate combinations of integrals and we have that

$$F_{10}E(X_1|X_1 > c, X_2 \le c) = \phi(c) - F_{11}E(X_1|X_1 > c, X_2 > c), \tag{7}$$

$$F_{01}E(X_1|X_1 \le c, X_2 > c) = \rho\phi(c) - F_{11}E(X_1|X_1 > c, X_2 > c), \tag{8}$$

$$\begin{aligned}F_{00}E(X_1|X_1 \le c, X_2 \le c) &= E(X_1) - F_{11}E(X_1|X_1 > c, X_2 > c) \\ &\quad - F_{01}E(X_1|X_1 \le c, X_2 > c) - F_{10}E(X_1|X_1 > c, X_2 \le c).\end{aligned} \tag{9}$$

As mentioned, we have no interaction if $\mu_{11} - \mu_{01} - \mu_{10} + \mu_{00} = 0$. Using Eqs. (2), (3), (4), (5), this leads to

$$\begin{aligned}&(\beta_1 + \beta_2)(E(X_1|X_1 > c, X_2 > c) + E(X_1|X_1 \le c, X_2 \le c) \\ &\quad - E(X_1|X_1 \le c, X_2 > c) - E(X_1|X_1 > c, X_2 \le c)) = 0\end{aligned} \tag{10}$$

From this, it is immediately clear that we can still have a spurious interaction even in situations where one of the two exposure variables is not associated with the outcome (β_1 or β_2 equal zero) as the product in (10) can still be different from zero. Furthermore, no spurious interaction can take place if $\beta_1 = -\beta_2$. However, this last point is a function of our highly symmetrical situation and hence less relevant.

Using Eqs. (6), (7), (8), (9) and the fact that in our example, $E(X_1) = 0$ and $F_{01} = F_{10}$, formula (10) can be written

$$(\beta_1+\beta_2)(1+\rho)\phi(c)\left[\frac{1-\Phi\left(\frac{c-\rho c}{\sqrt{1-\rho^2}}\right)}{F_{11}}-\frac{\Phi\left(\frac{c-\rho c}{\sqrt{1-\rho^2}}\right)}{F_{00}}-\frac{2\Phi\left(\frac{c-\rho c}{\sqrt{1-\rho^2}}\right)-1}{F_{01}}\right]=0 \quad (11)$$

We have to remember that F_{11}, F_{00} and F_{01} are also functions of c and ρ. Solving this (numerically) leads to $\rho = 0$ and / or $c = 0$, in addition to $\beta_1 = -\beta_2$ (and the degenerate solution $\rho = -1$). That is, we have no interaction if X_1 and X_2 are uncorrelated ($\rho = 0$) or if we split at the median ($c = 0$). Otherwise, categorization leads to spurious interaction. By the same arguments one can show that a spurious interaction will also appear if only one of the two variables X_1 and X_2 are categorized, or if X_1 and X_2 are split into more than two categories.

One should remember that Eq. (11) is only relevant for our specific situation, with normally distributed variables and common cut point c. The important message here is that the equation is true in only very few situations, which means that with a few exceptions, categorization leads to spurious interaction. This general message is true also for other distributions.

Having established the presence of spurious interaction, it is of interest to investigate the potential size of the problem. Based on the model above, we can investigate the size of the induced interaction term relative to main effects of the categorized versions of X_1, X_2. Assume we fit a model $Y = \tilde{\beta}_0 + \tilde{\beta}_1\tilde{X}_1 + \tilde{\beta}_2\tilde{X}_2 + \tilde{\beta}_3\tilde{X}_1\tilde{X}_2 + \tilde{\varepsilon}$ where $\tilde{X}_1, \tilde{X}_2$ are coded 0, 1. It is easy to show that $\tilde{\beta}_2$ is given by $\mu_{01} - \mu_{00}$ where μ_{00} and μ_{01} are given in (2) and (3), and we have already established that $\tilde{\beta}_3$ is given by $\mu_{11} - \mu_{01} - \mu_{10} + \mu_{00}$. We can use this to investigate the size of $\tilde{\beta}_3$ relative to $\tilde{\beta}_2$ for different choices of β_1, β_2, the cut point c and the correlation ρ. Figure 1 gives the absolute value of $\tilde{\beta}_3/\tilde{\beta}_2$ for varying c and ρ, for $\beta_1 = \beta_2 = 1$. We observe that with increasing cut points, the influence of the induced product term becomes substantial, even for moderate values of the correlation ρ. In this case, the ratio of $\tilde{\beta}_3$ to $\tilde{\beta}_1$ will of course be the same as to $\tilde{\beta}_2$. As a bi-product we can also study the size of $\tilde{\beta}_1$ and $\tilde{\beta}_2$ as a function of the same cut point c and correlation ρ (Fig. 2), and we observe that the estimated effects increase quite rapidly with the more extreme cut points and correlations.

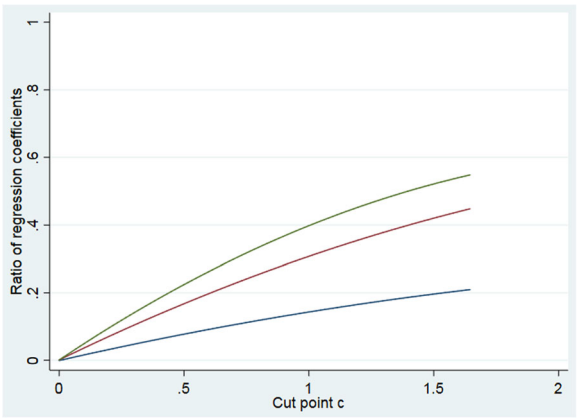

Fig. 1 The figure gives the absolute value of the ratio between $\tilde{\beta}_3$ and $\tilde{\beta}_i$, $i = 1, 2$ for $\rho = 0.2$ (blue line), 0.5 (red line) and 0.7 (green line) as a function of the cut point c when $\beta_1 = \beta_2 = 1$

Interpretation

Categorization of a continuous variable can be seen as an extreme form of measurement error. If there is measurement error in X_1 and / or X_2, and the error in X_1 is differential with respect to X_2 (meaning the measurement error in X_1 varies with X_2) or vice versa, it can be shown that this leads to an induced interaction between the observed versions of X_1, X_2 in a regression model [22–25]. In our case, the measurement error can be characterized by the reliability of the dichotomized variable $\tilde{X}_i$, relative to the continuous variable X_i, $i = 1, 2$, as measured by the point-biserial correlation. If X_i is normally distributed, this correlation is given by $h/\sqrt{pq}$, where h denotes the ordinate of the normal curve at the cut point and p and q denote the proportion of the population (or probability mass) above and below the cut point. It is easily seen that this correlation will vary with the other variable X_j, $i \neq j$, as p and q will vary with X_j. Hence, the measurement error is differential and an induced interaction is to be expected.

Another way of looking at the same problem is as a problem of residual confounding. If we have a situation with a confounding variable where the effect of the confounder is not properly adjusted for, we are left with

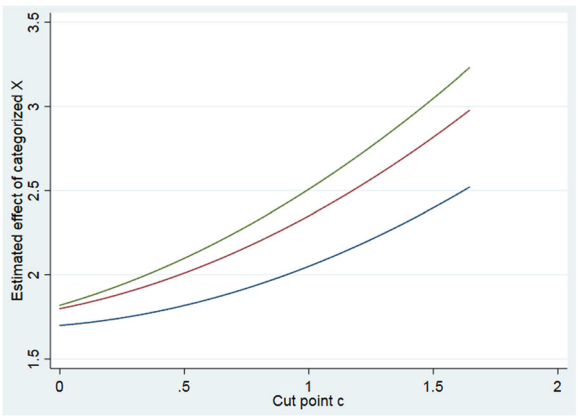

Fig. 2 The figure gives $\tilde{\beta}_2$ for $\rho = 0.2$ (blue line), 0.5 (red line) and 0.7 (green line) as a function of the cut point c. True $\beta_2 = \beta_1 = 1$

some residual confounding. This is exactly what is happening when a confounder is categorized. It is well known that this leads to biased estimates of exposure - outcome associations. In particular, Marshall and Hastrup [26] showed through simulations that such residual confounding can lead to apparent effects of variables that are strongly correlated to the confounder, but which in reality bear no association with the outcome. Marshall and Hastrup termed this effect "*resonance of strong confounders*". Important for our investigation, differences in residual confounding across strata may lead to spurious interaction [27].

Simulations

To illustrate our findings and to further explore the effects of dichotomization in different situations we conducted a small simulation study. Notice that our main focus here has been to investigate the qualitative aspects of the induced interactions, and hence, we used a rather large sample size to minimize randomness.

We simulated X_1, X_2 according to three different distributions; standard bivariate normal, uniform [0,1], and chi-square with 2 df. We let the correlation between X_1 and X_2, ρ, vary over 0.2, 0.5, and 0.7, and we categorized at the 60th and the 80th percentiles, respectively. Furthermore, we simulated a response Y according to the following model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ with ε normally distributed with zero expectation and variance σ^2 and independent of X_1, X_2. We then fit a model $Y = \tilde{\beta}_0 + \tilde{\beta}_1 \tilde{X}_1 + \tilde{\beta}_2 \tilde{X}_2 + \tilde{\beta}_3 \tilde{X}_1 \tilde{X}_2 + \tilde{\varepsilon}$ where $\tilde{\beta}_3$ is an interaction parameter. In all our simulations we let $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$. Furthermore, we let the residual variance σ^2 vary over the distributions in such a way that $\mathrm{Corr}(Y, X_i) = 0.3$, $i = 1, 2$, when $\rho = 0.5$. The correlation ρ was then varied without changing σ^2. The results of these simulations are given in Table 1. In addition, we repeated the situation with $\rho = 0.7$ and X_1, X_2 categorized at the 80th percentile, but now with $\beta_i = 2$ for $i = 1, 2$. Further, we simulated the same situation once more with $\beta_1 = 1$ and $\beta_2 = 0$. The results of these simulations are given in Table 2.

In order to generate correlated variates from the three distributions, we generated bivariate normal variates with a specified correlation structure in the standard way. Furthermore, we generated uniform marginals by applying the standard normal cumulative distribution function to each of the normal variates. Finally, on the basis of a uniform variate V_j, we can generate $X_j \sim$ chi-square with 2 df. by $X_j = -2 \ln(V_j)$ [28]. It is an easy task to adjust the pre-specified correlation structure so that the observed correlations are as one wish. This is done empirically, by running preliminary simulations. For each setting we ran 1000 simulations with a sample size of 10,000.

Table 1 Results of the simulation study

		Normal	Uniform	Chi-square
60th percentile $\rho = 0.2$	$\hat{\tilde{\beta}}_1$	1.74 (0.11)	0.54 (0.04)	2.99 (0.25)
	$\hat{\tilde{\beta}}_2$	1.73 (0.11)	0.54 (0.04)	3.00 (0.26)
	$\hat{\tilde{\beta}}_3$	−0.06 (0.20)	− 0.03 (0.06)	0.63 (0.40)
60th percentile $\rho = 0.5$	$\hat{\tilde{\beta}}_1$	1.89 (0.14)	0.59 (0.04)	2.81 (0.28)
	$\hat{\tilde{\beta}}_2$	1.88 (0.14)	0.59 (0.04)	2.83 (0.29)
	$\hat{\tilde{\beta}}_3$	−0.16 (0.21)	− 0.08 (0.06)	1.38 (0.43)
60th percentile $\rho = 0.7$	$\hat{\tilde{\beta}}_1$	1.95 (0.15)	0.62 (0.04)	2.60 (0.31)
	$\hat{\tilde{\beta}}_2$	1.94 (0.15)	0.61 (0.04)	2.63 (0.32)
	$\hat{\tilde{\beta}}_3$	−0.24 (0.22)	− 0.12 (0.07)	1.82 (0.48)
80th percentile $\rho = 0.2$	$\hat{\tilde{\beta}}_1$	1.96 (0.14)	0.57 (0.04)	4.22 (0.27)
	$\hat{\tilde{\beta}}_2$	1.96 (0.14)	0.57 (0.04)	4.23 (0.28)
	$\hat{\tilde{\beta}}_3$	−0.23 (0.26)	− 0.11 (0.07)	0.25 (0.55)
80th percentile $\rho = 0.5$	$\hat{\tilde{\beta}}_1$	2.22 (0.15)	0.67 (0.04)	4.35 (0.31)
	$\hat{\tilde{\beta}}_2$	2.22 (0.16)	0.67 (0.05)	4.35 (0.32)
	$\hat{\tilde{\beta}}_3$	−0.58 (0.25)	− 0.27 (0.07)	0.50 (0.54)
80th percentile $\rho = 0.7$	$\hat{\tilde{\beta}}_1$	2.35 (0.16)	0.74 (0.05)	4.35 (0.35)
	$\hat{\tilde{\beta}}_2$	2.36 (0.18)	0.74 (0.05)	4.34 (0.35)
	$\hat{\tilde{\beta}}_3$	−0.81 (0.26)	− 0.40 (0.08)	0.58 (0.56)

The true regression coefficients $\beta_i = 1$, $i = 1$, 2. The table gives estimated regression coefficients with corresponding empirical standard errors in parentheses

Results

Tables 1 and 2 give the results of these simulations. There are some common trends, but also some interesting differences between the distributions. First, the interaction effect becomes stronger with increasing correlation for all the distributions. We also observe that, naturally, the interaction effect becomes stronger with increasing main effects (the difference between Table 1 and the second situation in Table 2). In general, the normal distribution and the uniform distribution behave very similar. We

Table 2 Results of the simulation study

		Normal	Uniform	Chi-square
80th percentile $\rho = 0.7$ $\beta_1 = 1$, $\beta_2 = 0$	$\hat{\tilde{\beta}}_1$	1.65 (0.16)	0.50 (0.05)	3.35 (0.35)
	$\hat{\tilde{\beta}}_2$	0.70 (0.18)	0.24 (0.05)	0.99 (0.35)
	$\hat{\tilde{\beta}}_3$	−0.40 (0.26)	− 0.20 (0.08)	0.29 (0.55)
80th percentile $\rho = 0.7$ $\beta_i = 2$, $i = 1, 2$	$\hat{\tilde{\beta}}_1$	4.72 (0.17)	1.48 (0.05)	8.69 (0.36)
	$\hat{\tilde{\beta}}_2$	4.72 (0.18)	1.48 (0.05)	8.68 (0.37)
	$\hat{\tilde{\beta}}_3$	−1.63 (0.27)	− 0.79 (0.08)	1.16 (0.61)

The table gives estimated regression coefficients with corresponding empirical standard errors in parentheses

see a stronger interaction effect with the more extreme cut-off (80th percentile vs. 60th percentile). In the skewed chi-square distribution, we observe the opposite; the stronger interaction effects appear with a cut-off at the 60th percentile. Furthermore, for the normal- and uniform distributions, the estimated interaction parameter has a negative sign, while for the chi-square distribution it has a positive sign. Finally, we confirm the result from the theoretical calculations, that even in the situation with $\beta_2 = 0$, a spurious interaction appear.

If we look to the estimated standard errors, the interaction parameter becomes significant with a cut-off at the 60th percentile for the chi-square distribution in both Table 1 and Table 2, with an exception of the situation with the weaker correlation $\rho = 0.2$ in Table 1. For the other two distributions, we need in general to go to the more extreme cut-off (80th percentile) to find significant effects. Here, however, there are few significant effects for the chi-square distribution. It should be mentioned that categorization in general will lead to efficiency loss and the power to detect interaction effects will be low. For a general treatment of this topic, see [29].

To explain our findings, it will again be instructional to look at the problem as a measurement error problem. For the normal distribution, it is easy to show that the correlation between the continuous variable and the categorized version of the same variable is at its maximum with a cut-off at the median value, and it is then decreased with more extreme cut-off values. This means that there is more measurement error with the more extreme cut-off values. Based on the simulations, one can show that this is also true for the uniform distribution. However, also based on the simulations, for the chi-square distribution there is more measurement error with a cut-off at the 60th percentile than with a cut-off at the 80th percentile, as measured by the point-biserial correlation. This explains why we find the stronger interaction effects with a cut-off at the 60th percentile for the chi-square distribution, and with a cut-off at the 80th percentile for the two other distributions, as there will be more residual confounding, and hence more room for "resonance" in situations with more measurement error.

Furthermore, let us look into the differentiality of the measurement error. Thinking to the standard measurement error situation, with β_1 and $\beta_2 > 0$ we will have a positive interaction when there is less attenuation in the effect of X_1 for higher values of X_2 (and opposite), meaning the measurement error decreases with increasing X_2. Again based on the simulated data, we can look at the measurement error (the point-biserial correlation) of X_1 for $\tilde{X}_2 = 0$ vs. $\tilde{X}_2 = 1$. For the normal- and the Uniform distribution, we find more measurement error in X_1 for the lower category of $\tilde{X}_2$, while for the chi-square distribution, there is more measurement error in X_1 for the higher category of $\tilde{X}_2$. This explains why the sign of the interaction term differs between the distributions.

Collapsing categorical covariates

We will briefly mention another consequence of this type of induced interaction. It is well known that tests of interaction usually have rather low power. When modelling an interaction between categorical exposure variables that takes more than two categories, a number of extra parameters have to be introduced. A common advice is then to combine exposure groups in order to decrease the number of extra parameters and hence, increase power (see e.g. Kirkwood & Sterne, Ch. 29.5) [30]. However, by doing this we will again run the risk of introducing a spurious interaction, which can be easily realized by the following illustration.

Assume we have one binary exposure variable X_1 (e.g. gender) and one categorical exposure variable X_2 with four categories. These two exposure variables are to be related to a binary outcome (diseased / not diseased). Assume the data look as in Table 3.

As seen from the table, there is no interaction between X_1 and X_2, as the association between X_2 and the outcome as measured by the relative risk (RR) is constant across the two levels of X_1. Next, we will combine categories 1 and 2, and 3 and 4 of X_2, producing a new variable $\tilde{X}_2$ taking only two categories. This leads to Table 4.

As observed, the association between $\tilde{X}_2$ and the outcome now differs between the two levels of X_1 and a spurious interaction is introduced. Although the effect is not

Table 3 Illustration of the effect of collapsing exposure categories

X_1=1					X_1=2				
X_2	Diseased	Not diseased	Total	RR	X_2	Diseased	Not diseased	Total	RR
1	20	380	400	1.0	1	5	95	100	1.0
2	30	270	300	2.0	2	20	180	200	2.0
3	30	170	200	3.0	3	45	255	300	3.0
4	20	80	100	4.0	4	80	320	400	4.0

The table gives the true situation

Table 4 Illustration of the effect of collapsing exposure categories

X_1=1					X_1=2				
$\tilde{X}_2$	Diseased	Not diseased	Total	RR	$\tilde{X}_2$	Diseased	Not diseased	Total	RR
1	50	650	700	1.00	1	25	275	300	1.00
2	50	250	300	2.33	2	125	575	700	2.14

The table gives the observed situation after collapsing

particularly strong in this example, an apparent increase in power may partially be due to such an induced interaction.

Discussion

We have given yet another argument to why continuous explanatory variables should not be categorized when entered into a regression model. If we have correlated exposure variables and categorize, this may lead to spurious interactions in the regression model. Furthermore, we have given an interpretation of this as a measurement error problem.

Statistical interaction is scale dependent, both with regard to model and measurement. An additive effect on a linear scale may appear as multiplicative on a transformed scale, like the logit. In the same way, any monotone transformation of the measurement scales of X_1, X_2 may lead to interactions. Hence, statistical interactions need to be interpreted relative to the scales on which they appear. As such, the main problem with the type of interaction discussed in the present work is not its existence but the fact that it is typically interpreted relative to its original measurement scale.

Our formal development has been within the framework of the linear regression model. However, based on the considerations above, it is easy to realize that this holds also for other regression models. Indeed, we have shown the appearance of such an interaction effect in logistic models through an example.

The practical implication of this is that whenever an interaction effect appears in an analysis based on categorized explanatory variables, the categorization itself should be considered as a possible explanation. However, one should also be aware that such an induced interaction may counteract any possible true interaction present in the data, and hence, mask this true interaction. It should be mentioned that in a practical data analysis, one will need a rather large sample size or strong effects for these interactions to appear as statistically significant.

Conclusion

In summary, categorization of continuous variables should be avoided. It leads to a number of problems, including biased estimates, loss of power, inflated type-I error, and spurious interaction effects. If the true effect of the exposure variable(s) in question cannot be easily modelled by classical parametric models, non-parametric regression methods should be preferred in order to avoid the above-mentioned problems and to gain insight into the underlying relationship. If one choose to categorize despite such warnings, it is generally preferable to categorize into more than two groups in order to minimize the information loss.

Abbreviations

ANOVA: Analysis of variance; Corr: Correlation; Df: Degrees of freedom; Exp(.): The exponential function; l/min: Liters per minute; SE: Standard error

Acknowledgements

The author wants to thank the Department of Molecular Medicine, Division of Physiology at the University of Oslo, for providing the PEF data for the first example, and the Norwegian Institute of Public Health and the Norwegian Cause of Death Registry for providing the health survey data for the second example. The author also wants to thank PhD Ingvild Dalen for pointing him to the problem in the first place.

Funding

Not applicable.

Availability of data and materials

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. Code for the simulations was written in the Gauss software and can be obtained from the author.

Authors' contributions

The author read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares that he has no competing interest.

Publisher's Note

Received: 6 February 2018 Accepted: 21 January 2019
Published online: 07 February 2019

References

1. Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analysis. Stat Med. 2004;23:1159–78.
2. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med. 2006;25:127–41.
3. Altman DG, Royston P. The cost of dichotomizing continuous variables. BMJ. 2006;332:1080.
4. Frøslie KF, Røislien J, Laake P, et al. Categorisation of continuous exposure variables revisited. A response to the Hyperglycaemia and

adverse pregnancy outcome (HAPO) study. BMC Med Res Methodol. 2010;10:103.
5. Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. BMC Med Res Methodol. 2012;12:21.
6. Barnwell-Ménard J-L, Li Q, Cohen AA. Effects of categorization method, regression type, and variable distribution on the inflation of type-I error rate when categorizing a confounding variable. Stat Med. 2015;34:936–49.
7. Collins GS, Ogundimu EO, Cook JA, et al. Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model. Stat Med. 2016;35:4124–35.
8. Weinberg CR. How bad is categorization? Epidemiology. 1995;6:345–7.
9. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. Epidemiology. 1995;6:356–65.
10. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology. 1995;6:451–4.
11. Brenner H, Blettner M. Controlling for continuous confounders in epidemiologic research. Epidemiology. 1997;8:429–34.
12. Van Walraven C, Hart RG. Leave'em alone – why continuous variables should be analyzed as such. Neuroepidemiology. 2008;30:138–9.
13. Schellingerhout JM, Heymans MW, de Vet HCW, et al. Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain. J Clin Epidemiol. 2009;62:868–74.
14. Maxwell SE, Delaney HD. Bivariate median splits and spurious statistical significance. Psychol Bull. 1993;113:181–90.
15. Vargha A, Rudas T, Delaney HD, et al. Dichotomization, partial correlation, and conditional independence. J Educ Behav Stat. 1993;21:264–82.
16. MacCallum RC, Zhang S, Preacher KJ, et al. On the practice of dichotomization of quantitative variables. Psychol Methods. 2002;7:19–40.
17. Humphreys LG, Fleishman A. Pseudo-orthogonal and other analysis of variance designs involving individual-differences variables. J Educ Psychol. 1974;66:464–72.
18. Paunonen SV, Jackson DN. Idiographic measurement strategies for personality and prediction: some unredeemed promissory notes. Psychol Rev. 1985;92:486–511.
19. Bissonnette V, Ickes W, Berstein E. Personality moderating variables: a warning about statistical artifact and a comparison of analytic techniques. J Pers. 1990;58:567–87.
20. Regier MH, Hamdan MA. Correlation in a bivariate normal distribution with truncation in both variables. Aust J Stat. 1971;13:77–82.
21. Mehler FG. Uber die Entwicklung einer Funktion von beliebig vielen Variablen nach Laplaceschen Funktionen höherer Ordnung. J Fur Mathematik. 1866;66:161–76.
22. Greenland S. The effect of misclassification in the presence of covariates. Am J Epi. 1980;112:564–9.
23. Greenland S. Basic problems in interaction assessment. Environ Health Perspect. 1993;101:59–66.
24. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. Occup Environ Med. 1998;55:651–6.
25. Muff S, Keller LF. Reverse attenuation in interaction terms due to covariate measurement error. Biom J. 2015;57:1068–83.
26. Marshall JR, Hastrup JL. Mismeasurement and the resonance of strong confounders: uncorrelated errors. Am J Epi. 1996;143:1069–78.
27. Pearce N, Confounding GS. Interaction. In: Handbook of epidemiology. 2nd ed. New York: Springer; 2014.
28. Johnson NL, Kotz S, Balakrishnan N. Continuous Distributions, Vol. 1, 2nd ed (p. 347). New York: Wiley; 1994.
29. Farewell VT, Tom BDM, Royston P. The impact of dichotomization on the efficiency of testing for an interaction effect in exponential family models. J Am Stat Assoc. 2004;99:822–31.
30. Kirkwood BR, Sterne JAC. Essential medical statistics. 2nd ed. Oxford: Wiley-Blackwell; 2003.