Effects of categorization method, regression type, and variable distribution on the inflation of Type-I error rate when categorizing a confounding variable

Running head: Categorized confounders and Type-I error

Jean-Louis Barnwell-Ménard[1], Qing Li[2], and Alan A. Cohen[2*]

[1]Department of Economics, University of Sherbrooke, Sherbrooke, QC, Canada

[2]Department of Family Medicine, University of Sherbrooke, Sherbrooke, QC, Canada

*Corresponding author: Groupe de recherche PRIMUS, Faculté de Médecine et Sciences de la Santé, CHUS-Fleurimont, 3001 12^{e} Ave Nord, Sherbrooke, QC J1H 5N5 Canada Alan.Cohen@USherbrooke.ca, (819) 346-1110 x12589, (819) 820-6419 (fax)

Abstract

The loss of signal associated with categorizing a continuous variable is well known, and previous studies have demonstrated that this can lead to an inflation of Type-I error when the categorized variable is a confounder in a regression analysis estimating the effect of an exposure on an outcome. However, it is not known how the Type-I error may vary under different circumstances, including logistic versus linear regression, different distributions of the confounder, and different categorization methods. Here we analytically quantified the effect of categorization, and then performed a series of 9600 Monte Carlo simulations to estimate the Type-I error inflation associated with categorization of a confounder under different regression scenarios. We show that Type-I error is unacceptably high (>10% in most scenarios, and often 100%). The only exception was when the variable categorized was a continuous mixture proxy for a genuinely dichotomous latent variable, where both the continuous proxy and the categorized variable are error-ridden proxies for the dichotomous latent variable. As expected, error inflation was also higher with larger sample size, fewer categories, and stronger associations between the confounder and the exposure or outcome. We provide online tools that can help researchers estimate the potential error inflation and understand how serious a problem this is.

Keywords: Type-I error, confounding, categorization, dichotomization, simulation, distribution

Introduction

Researchers and clinicians in epidemiologic and medical studies often categorize continuous variables for purposes of facilitating the interpretability of results [1] (common examples include age, body-mass index, socio-economic status, and levels of blood biomarkers). The unnecessary use of categorical variables has been criticized by many for the potential increase in statistical bias and the loss of information [2-17], but use of categorized continuous variables is still standard practice in the epidemiologic and medical literature [18]. There is a consensus among statisticians that statistical tools treating variables as continuous (e.g. with non-parametric or spline regressions) are preferred and more robust when the latent trend is not easily captured by classical parametric models [2, 7, 17]. Such tools are, however, more complex to apply and interpret for clinicians, which might be a reason for the continued abundant use of categorized data in epidemiological publications.

A specific situation prominent in epidemiologic and medical research where categorized continuous variables are regularly used is for control variables (confounding variables) in regression models when assessing the potential impact of an exposure (risk factor, independent variable of interest) on an outcome (dependent variable). Confounding variables are defined here as variables that are associated with both the exposure of interest and the outcome of interest, but which are not affected by either variable [19]. Unlike a categorization of the exposure or outcome variables, which can lead to an inflation of Type-II error [20], categorization of a confounding variable can lead to increased residual confounding, i.e., effects of confounding variables that are unmeasured and thus not accounted for in the model. Such residual confounding generally results in the detection of spurious relationships between the exposure and the outcome, and thus false rejection of null hypotheses (inflated Type-I error) because the model does not replicate perfectly the statistical relationship between the confounding variable and the concerned variables in models [17]. Austin & Brunner [7] assessed the influence such methodology has on the statistical performance of models under the hypothesis of normal variable distributions and logistic regression. They demonstrated important residual confounding sufficient to suggest that researchers may often falsely detect a potential association between an exposure and an outcome.

Quantiles and clinical cut-offs are the most common methods for categorizing continuous confounding variables [18]. Clinicians and epidemiologists frequently study variables with various distribution shapes and select their cut-offs (i.e., through a categorization method) in order to minimize the loss of information or to group similar observations. In spite of the common categorization methodologies, little is known about how cut-off selection, variable distributions, or type of regression model (linear, logistic) might affect the statistical bias and robustness of the results induced by the categorization of confounding variables.

Because unnecessary categorization is such a rampant problem, it is important to understand what factors contribute to greater error inflation when categorizing, and to quantify error inflation under different scenarios. The ability to quantify error inflation could become a tool to force researchers to consider more carefully the consequences of categorization on their conclusions. In this paper, we assessed how generalizable the conclusions of Austin & Brunner [7] were across a wide range of realistic data analysis scenarios, and whether there might be some cases where the implications of categorization were particularly severe. We simulated the rate of falsely rejecting the true value of the coefficient relating an exposure to an outcome (the Type-I error) under different scenarios where a confounding variable is categorized. In addition, we mathematically show the effect of categorization for the case of linear regression. We have also developed a statistical application available on the web allowing easy estimation of the Type-I error rate under different categorization algorithms for varying statistical hypotheses.

Mathematical Derivation

The categorization of a confounding variable generates measurement error with respect to the original variable. We recapitulate this effect with the following mathematical derivation in the case of linear regression because it is possible to get a closed-form expression of the asymptotic bias which allows seeing immediately the determinants. The literature origin of the effect is well exposed in [21], as well as the risk for measurement error in general for different error sources and regression scenarios. Under these circumstances the estimators are asymptotically biased, affecting the estimated values, the confidence intervals and consequently the Type-1 error rate.

For individual *i* the model is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

The confounding variable x_{2i} is categorized into

$$x_{2i}^c = x_{2i} + u_i$$

where the superscript “c” denotes “categorical”. Under the assumption that $\mathrm{E}(x_{1i}x_{2i}) \neq 0$, and since the value of x_{2i} decides which category the individual *i* goes into, we know that $cov(x_{2i}, u_i) \neq 0$ and hence $cov(x_{1i}, u_i) \neq 0$. The term u_i is the difference between x_{2i}^c and x_{2i} for individual i, i.e. the measurement error introduced by categorization. Note that $E(u_i) \neq 0$, and in addition, the measurement error u_i is correlated with the true value x_{2i} which is different from classical measurement error models; this case was discussed in [22] and the correlation between u_i and x_{2i} has an influence on the analytical expression of the bias, making the bias more unpredictable. We make the classical assumptions of orthogonality for linear regression, i.e. $E(x_{1i}\varepsilon_i) = 0$ and $E(x_{2i}\varepsilon_i) = 0$, which leads to $E(u_i\varepsilon_i) = 0$ and $E(x_{2i}^c\varepsilon_i) = 0$. Plugging x_{2i}^c into the regression gives

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}^c + \varepsilon_i - \beta_2 u_i$$

Letting $v_i = \varepsilon_i - \beta_2 u_i$, we get $cov(x_{1i}, v_i) = -\beta_1\beta_2 cov(x_{1i}, u_i) \neq 0$ and $cov(x_{2i}^c, v_i) = -\beta_2^2 cov(x_{2i}^c, u_i) \neq 0$. In matrix form, defining

$$\beta \coloneqq \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \; y \coloneqq \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \; v \coloneqq \begin{bmatrix} \varepsilon_1 - \beta_2 u_1 \\ \vdots \\ \varepsilon_N - \beta_2 u_N \end{bmatrix} \; X \coloneqq \begin{bmatrix} 1 & x_{11} & x_{21}^c \\ \vdots & \vdots & \vdots \\ 1 & x_{1N} & x_{2N}^c \end{bmatrix}$$

for a sample with size N, we can write the regression as $y = X\beta + v$. Hence, the classical least squares estimator $\hat{\beta}$ converges asymptotically to

$$\operatorname*{plim}_{N\to\infty} \hat{\beta} = \operatorname*{plim}_{N\to\infty} (X'X)^{-1} X'y = \beta + \operatorname*{plim}_{N\to\infty} (X'X)^{-1} X'v$$

and the asymptotic bias generated by categorization (or by introducing measurement error, in a broader sense) is

$$bias(\hat{\beta}) = \operatorname*{plim}_{N\to\infty} (X'X)^{-1} X'v = \operatorname*{plim}_{N\to\infty} \left(\frac{1}{N} X'X\right)^{-1} \operatorname*{plim}_{N\to\infty} \frac{1}{N} X'v$$

Since,

$$\operatorname*{plim}_{N\to\infty}\left(\frac{1}{N}X'X\right)^{-1}=\operatorname*{plim}_{N\to\infty}\begin{bmatrix}1 & \frac{\sum_{i=1}^{N}x_{1i}}{N} & \frac{\sum_{i=1}^{N}x_{2i}^{c}}{N}\\ \frac{\sum_{i=1}^{N}x_{1i}}{N} & \frac{\sum_{i=1}^{N}x_{1i}^{2}}{N} & \frac{\sum_{i=1}^{N}x_{1i}\,x_{2i}^{c}}{N}\\ \frac{\sum_{i=1}^{N}x_{2i}^{c}}{N} & \frac{\sum_{i=1}^{N}x_{1i}\,x_{2i}^{c}}{N} & \frac{\sum_{i=1}^{N}x_{2i}^{c2}}{N}\end{bmatrix}^{-1}$$

and

$$\operatorname*{plim}_{N\to\infty}\frac{1}{N}X'v=\operatorname*{plim}_{N\to\infty}\begin{bmatrix}\frac{\sum_{i=1}^{N}(\varepsilon_i-\beta_2u_i)}{N}\\ \frac{\sum_{i=1}^{N}[x_{1i}(\varepsilon_i-\beta_2u_i)]}{N}\\ \frac{\sum_{i=1}^{N}[x_{2i}^{c}(\varepsilon_i-\beta_2u_i)]}{N}\end{bmatrix}$$

using Slutsky theorem and the property $\operatorname*{plim}_{N\to\infty}\frac{\sum_{i=1}^{N}x_i}{N}=E(x_i)$, we get

$$bias(\hat{\beta})=\operatorname*{plim}_{N\to\infty}\left(\frac{1}{N}X'X\right)^{-1}\operatorname*{plim}_{N\to\infty}\frac{1}{N}X'v$$

$$=\begin{bmatrix}1 & E(x_{1i}) & E(x_{2i}^{c})\\ E(x_{1i}) & E(x_{1i}^{2}) & E(x_{1i}x_{2i}^{c})\\ E(x_{2i}^{c}) & E(x_{1i}x_{2i}^{c}) & E(x_{2i}^{c2})\end{bmatrix}^{-1}\begin{bmatrix}E(\varepsilon_i)-\beta_2E(u_i)\\ E(x_{1i}\varepsilon_i)-\beta_2E(x_{1i}u_i)\\ E(x_{2i}^{c}\varepsilon_i)-\beta_2E(x_{2i}^{c}u_i)\end{bmatrix}$$

The last matrix product leads to a 3 × 1 matrix where the three elements correspond to the asymptotic biases of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. With the assumptions $E(\varepsilon_i)=E(x_{1i}\varepsilon_i)=E(x_{2i}\varepsilon_i)=0$ and some basic calculations we get the following expression for the second element of the matrix, namely $bias(\hat{\beta}_1)$ which is equal to

$$\frac{\beta_2[E(u_i)E(x_{1i}x_{2i}^{c})E(x_{2i}^{c})-E(x_{1i}u_i)E^2(x_{2i}^{c})+E(x_{1i}u_i)E(x_{2i}^{c2})-E(u_i)E(x_{1i})E(x_{2i}^{c2})+E(x_{1i})E(x_{2i}^{c})E(x_{2i}^{c}u_i)-E(x_{1i}x_{2i}^{c})E(x_{2i}^{c}u_i)]}{E^2(x_{1i}x_{2i}^{c})-2E(x_{1i})E(x_{2i}^{c})E(x_{1i}x_{2i}^{c})+E(x_{1i}^{2})\left(E^2(x_{2i}^{c})-E(x_{2i}^{c2})\right)+E^2(x_{1i})E(x_{2i}^{c2})}$$

The last expression, which shares similarities to the bias expression found by [21], finds that the asymptotic bias of $\hat{\beta}_1$ depends on the value of β_2, but does not depend on the value of β_1 itself. Also, the bias is affected by the first and second order moments related to x_{2i}^{c} and u_i which depend on the method of categorization as well as the distributions of the original variables. The analytical expression is non-linear in the relevant moments and so it is not easy to characterize the effect of a single determinant (e.g. method of categorization, data distribution, number of categories, etc.); in practice, the expression will become even more complex when adding additional regressors, but in a word, it is the introduction of measurement error that creates the bias, whatever the nature of the original variables is. Importantly, the complexity of this expression shows that the precise magnitude of the bias is not easily predictable. Simulations in the following sections give intuitive results in different cases.

Methods

Our simulations were modeled on the approach of Austin & Brunner [7]. We simulated data under the general scenario of the following regression model:

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \nu$$

where Y is an outcome of interest, X_1 is an exposure whose relationship to Y we would like to assess, and X_2 is a potential confounding variable which is available in continuous format but which is categorized for analysis. The true values of β_0 and β_1 are assumed to be zero (i.e., X_1 has no direct effect on Y, since we wish to evaluate the Type-I error), and β_2 has a specified positive value. Parameters which were allowed to vary included (a) type of regression model (linear versus logistic), (b) distribution of the underlying confounder (X_2), (c) the covariance between X_1 and X_2, (d) β_2, (e) the method for categorizing X_2 when continuous, (f) the number of categories into which X_2 is divided, and (g) the sample size of the simulated data set.

Data generation

Continuous confounding variable

An exposure (X_1, assumed to be independent of the outcome) and a continuous confounding variable (X_2) were generated with three different processes in order to assess the confounding variable under (1) normal, (2) log-normal or (3) bimodal distributions.

(1) The first process ("normal") for the generation of a normal exposure $X_{1,n}$ and a normal confounding variable $X_{2,n}$ used a bivariate normal distribution of size N with mean $\mu = (0,0)$ and covariance matrix $\Omega = \begin{pmatrix} 1 & \sigma_{1,2} \\ \sigma_{1,2} & 1 \end{pmatrix}$ where $\sigma_{1,2}$, the confounder-exposure covariance, ranged from 0 to 0.9 in increments of 0.1.

(2) The second process ("log-normal") for the generation of a normal exposure $X_{1,l}$ and a log-normal confounding variable $X_{2,l}$ was obtained with the exponential transformation of a normal confounding variable $X_{2,n}$ generated with process (1). The average sampled kurtosis of $X_{2,l}$, out of 1000 samples with arbitrary covariance specification and sample size of 2000, was 62.04, with a 95% bootstrap confidence interval for the sample kurtosis average ranging from 56.88 to 67.18, and the average skewness was 5.22, with a 95% bootstrap confidence interval for the sample skewness average ranging from 5.10 to 5.35.

(3) The third process ("bimodal") for the generation of a normal exposure $X_{1,b}$ and a potentially correlated bimodal confounding variable $X_{2,b}$ was based on the separate simulation of two groups of data, *I* and *II*, representing each of the modes in $X_{2,b}$ (i.e., $X^1_{2,b}$ and $X^2_{2,b}$) along with their paired values in $X_{1,b}$ (i.e.,

$X_{1,b}^1$ and $X_{1,b}^2$). $X_{1,b}^1$ and $X_{2,b}^1$ (I) were simulated from a bivariate normal distribution of size N_1 with mean $\mu_1 = (0, 0)$ and covariance matrix $\Omega_1 = \begin{pmatrix} 1 & \sigma_{1,2} \\ \sigma_{1,2} & 1 \end{pmatrix}$. $X_{1,b}^2$ and $X_{2,b}^2$ (II) were simulated from a bivariate normal distribution of size N_2 with mean $\mu_2 = (0, U(3,4))$ and covariance matrix $\Omega_2 = \begin{pmatrix} 1 & \sigma_{1,2} \\ \sigma_{1,2} & U(4,9) \end{pmatrix}$. Once the four variables were simulated, $X_{1,b}$ was generated as the union of $X_{1,b}^1$ and $X_{1,b}^2$, and $X_{2,b}$ was generated as the union of $X_{2,b}^1$ and $X_{2,b}^2$, keeping their relative orders so as to maintain the pairing of values and thus the correlation. $\sigma_{1,2}$ ranged from 0 to 0.9 by increments of 0.1. U represents the uniform distribution (e.g., min = 3 and max = 4). Total sample size $N = N_1 + N_2$, but $N_1 \neq N_2$. This simulation method allowed $X_{1,b}$ and $X_{2,b}$ to covary at level $\sigma_{1,2}$ even while $X_{1,b}$ represents a unimodal normal distribution and $X_{2,b}$ represents a bimodal distribution generated as a mixture of two normal distributions with different means and variances. The average sampled kurtosis of $X_{2,b}$, out of 1000 samples with arbitrary covariance specification and sample size of 2000, was 5.29, with a 95% bootstrap confidence interval for the sample kurtosis average ranging from 5.25 to 5.33, and the average skewness was 1.45, with a 95% bootstrap confidence interval for the sample skewness average ranging from 1.44 to 1.46.

Proxy variable for a dichotomous underlying confounder

In addition to the three above scenarios featuring continuous confounding variables with different distributions, we simulated a fourth scenario in which the true confounding variable is dichotomous but researchers only observe a continuous proxy. This corresponds in reality to using blood glucose level as a continuous proxy for underlying diabetes state, or to using a sex steroid level to assign sex when true sex is unknown. If the true confounder is the underlying dichotomous variable, we might ask to what extent we can categorize the proxy in order to better approach the true confounder (supposing that it is known that proxy is not the true confounder). The exposure (X_1), the proxy confounding variable (X_2), and the underlying dichotomous confounding variable (X_3) were generated with the following fourth process ("mixture distribution"):

(4) $X_{1,d}$ (the normal exposure) and $X_{2,d}$ (the bimodal proxy confounding variable) were generated identically as in process (3), the mixture of two multivariate normal distributions (I) and (II) of size $N = N_1 + N_2$. $X_{3,d}$ (the underlying dichotomous confounding variable) is a dummy variable taking the following values:

$$\begin{cases} if\ X_{2i,d} \in (I):\ X_{3i,d} = 0 \\ if\ X_{2i,d} \in (II):\ X_{3i,d} = 1 \end{cases}$$

Outcome variable (continuous confounder)

Once the unrelated exposure $X_{1,(n,l\ or\ b)}$ and the confounding variable $X_{2,(n,l\ or\ b)}$ were generated, the outcome (independent) variable $Y_{(n,l\ or\ b)}$ could be obtained using (a) a logistic model or (b) a linear model for its generation in the following procedure:

(a) Logistic model:

$$logit(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \text{for } i = 1, 2, \dots, N$$

$$\text{where, } p_i = \frac{\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i})}{\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i})+1}, \text{for } i = 1, 2, \dots, N$$

$$y_i \sim Binomial(p_i), \text{ for } i = 1, 2, \dots, N$$

X_1 denotes the *de facto* unrelated exposure and X_2 is the confounding variable correlated with X_1 and the outcome Y. The logistic model was assessed for five confounder-outcome association scenarios:

$$\beta_0 = 0, \beta_1 = 0 \text{ and } \beta_2 = (0.2, 0.5, 1, 2, 3)$$

where the range for the predetermined values of β_2 was based upon Austin & Brunner [7] modeling scenarios, with the addition of 0.2 and 2 for generality purposes.

(b) The linear model:

$$y_i = x_{2i} + \varepsilon_{i,} \text{ for } i = 1, 2, \dots, N$$

where x_{2i} is treated as a constant and $\varepsilon \sim N(0, \sigma^2)$. Therefore, as σ^2 increases, we expect a lower predictive power of the outcome variable (y) by the confounding variable (x_2), which correspond to the idea of a decreasing value of β_2 in the logistic model. The linear model was assessed for five confounder-outcome association scenarios:

$$\sigma^2 = (9.95, 3.17, 1.73, 1.02, 0.48)$$

The values for σ^2 were chosen empirically via simulations to correspond as closely as possible to values of β_2 for a residual confounding effect equivalent to those used in (a) for the logistic model.

Outcome variable (dichotomous underlying confounder)

Once the unrelated exposure $X_{1,d}$, the bimodally distributed proxy representing the dichotomous confounder $X_{2,d}$ and the underlying dichotomous confounder $X_{3,d}$ were generated, the outcome (independent) variable Y_d could be obtained using both models

(a) and (b), with the sole difference here that X_2 is replaced by X_3 in the generating procedure. Therefore, the logistic and linear model become, respectively:

(a)

$$logit(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{3i}, \text{ for } i = 1, 2, \ldots, N$$

$$where, p_i = \frac{\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{3i})}{\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{3i})+1}, \text{for } i = 1, 2, \ldots, N$$

$$y_i \sim Binomial(p_i), \text{ for } i = 1, 2, \ldots, N$$

(b)

$$y_i = x_{3i} + \varepsilon_{i,} \text{ for } i = 1, 2, \ldots, N$$

Both models use the same confounder-outcome association scenarios as with the continuous confounding variable modeling. $X_{3,d}$ is used only to generate Y_d; once Y_d is generated, the dichotomous variable $X_{3,d}$ is represented by its proxy variable $X_{2,d}$ (bimodally distributed) in the model estimating the Type-I error rates. The mixture and bimodal distributions thus differ only in that the outcome is determined directly by the continuous bimodal confounder in the bimodal distribution, but is determined by the underlying dichotomous variable in the mixture distribution.

Categorization algorithms

The Type-I error for the true null hypothesis of the unrelated exposure was assessed with the confounding or proxy variable categorized in two, three, four and five categories, or kept continuous for comparison. The confounding variable was categorized using two different methods: (A) quantile and (B) maximized R^2.

(A) The first method consists in dividing the confounding variable into quantiles, i.e. separating the sorted x_{2i}, for $i = 1, 2, \ldots, N$, in groups with an equal number of observations. This method is arguably the most frequently used in practice, and was explained in detail by [7].

(B) The second method finds category cut-offs that optimize the linear fit of a continuous variable by the same categorized variable. The optimal cut-offs define the categories that maximize the adjusted R^2 of the following preliminary linear model (which differs from models (1) and (2)):

$$X_2 = \alpha_0 + X_2^c \alpha_1 + \mu$$

where X_2 corresponds to the continuous variable and X_2^c to the same categorized variable. The optimal cut-offs are found using a linear optimization function for a one cut-off search and a non-linear optimization function for a 2-4 cut-off search (with the optimize and optim functions in R). We applied this method with 1, 2, 3 and 4 cut-offs, giving a categorized confounding variable (X_2^c) with two, three, four and five categories respectively.

Simulations of Type I error

Using the framework above, we had eight independent parameters that could be adjusted in the simulations: (1) Underlying confounder distribution (4 levels: normal, log-normal, bimodal, or dichotomous); (2) Regression type (2 levels: logistic or linear); (3) Categorization method (2 levels: quantile or maximized R^2); (4) Category number (4 levels: 2-5); (5) Confounder-exposure covariance (10 levels: 0 - 0.9 in increments of 0.1); (6) Confounder-outcome association (5 levels: β_2 or σ^2); (7) Sample size of the simulated study (3 levels: 100, 500 or 2000); (8) Number of Monte Carlo iterations per scenario (1 level used: 1000 iterations). Monte Carlo simulations were performed for all

9600 combinations of these parameters. For each parameter combination, we calculated the Type-I error rate as the percentage of the 1000 Monte Carlo iterations in which the *p*-value of the following parameter significance t-test:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

was less than α=0.05, i.e. falsely rejecting the true null hypothesis of no relationship between the exposure and the outcome with a confidence level of 95%.

Summarizing results

Because of the large number of results generated by these simulations, we used three parallel methods to summarize our results. First, we conducted linear regression models on the database of simulation results, modeling the Type-I error rate among the thousand iterations as a function of the seven varying parameters included in the models. We also stratified and included interactions as necessary. Presentation of results is stratified between the normal, log-normal and bimodal confounder distributions on the one hand and the mixture distribution on the other, given that the latter is a special case with particular properties. In order to show the approximate magnitude of effects, we present results of regression models as if effects were linear and additive (e.g., change in Type-I error for each change of 0.1 in σ), though clearly this is not strictly true and should not be taken overly literally. Second, we developed an online interactive interface that allows users to choose parameters of interest and generate figures similar to those shown here in order to graphically examine several parameters and their interactions, https://usherbrookeprimus.shinyapps.io/resultsApp/. Third, we present a selection of results from the online tool as figures to illustrate key points.

Results

Performance of categorization methods

For a normally distributed confounding variable, the quantile and maximized R^2 methods provided essentially identical categories. For a log-normally distributed confounding variable, the maximized R^2 method provided cut-offs that were substantially further toward the tail of the distribution than those chosen by the quantile method. For the bimodal distribution ($X_{2,b}$ or $X_{2,d}$), the maximized R^2 method was substantially better at separating the two modes near the bottom of the trough (Figure 1), especially with only 2 categories (referred to hereafter as "optimal categorization").

Type-I error: Normal, log-normal, and bimodal confounder distributions

The results from our simulations demonstrated a substantial inflation of the Type-I error rate for detecting an effect of the unrelated exposure (X_1) on the outcome (Y) when the confounder (X_2) was categorized, except when the confounder was very weakly associated with either the exposure or the outcome (Table 1, top). As expected, Type-I error rate always increased as the correlation between the exposure and the confounder (σ) increased, with approximately 5.6% additional error for each increment of 0.1 in σ (Figure 2). Type-I error rate decreased monotonically as the number of categories increased, with approximately 7.9% fewer errors for each additional category added

(Figure 2). Accordingly, a confounder categorized in five categories obtained a lower Type-I error rate compared to a confounder with two, three and four categories, with the exception of the bimodal distribution (3) categorized with the quantile method (A) under the linear model (a), where 3 categories minimized the type one error rate. Each additional 100 added to sample size increased the Type-I error by about 0.96%, or about 14.4% higher rates with sample size =2000 than =500 (Figure 3). Additionally, there was about a 8.8% increase in Type-I error for each additional increment of association between the confounder (X_2) and the outcome (Y) (Figure 3). The quantile categorization method obtained lower Type-one error rates (Figure 4) for the three distributions. The distribution type did not express a clear pattern for minimizing the error rate.

In sum, under all scenarios, with the exception of a very weak confounder-outcome or confounder-exposure association (where the addition of a confounding variable is not as relevant), categorizing a continuous confounding variable substantially inflated the risk for type-I error rate. Although it might seem intuitive to dichotomize a bimodal confounder, we found that with the bimodal confounder distribution (3) and the maximized R^2 categorization method (B) even an "optimal" categorization process significantly inflated the type-I error rate, performing even worse than an arbitrary categorization criterion such as with the quantile method.

Type-I error: Dichotomous unmeasured confounder (mixture distribution)

With a dichotomous unmeasured confounder represented by a bimodal continuous proxy, results were qualitatively similar to results under other distributions for sample size, the strength of the confounder-exposure correlation (σ), and the strength of the confounder-outcome association (β_2 or σ^2), and are not discussed further (Table 1, bottom). However, inversed results were found for the number of categories and the categorization method on the proxy variable (Figure 5). Two categories with the maximized R^2 method now performed best, with worse results for three (4.5% more error), four-five categories (6% more error), and the quantile method in general. The dichotomized proxy confounder gave lower Type-I error rates than its continuous state, although its error rates were still substantial. The maximized R^2 method performed better, with a 10% lower error rate, though this effect was attenuated substantially with more than two categories: by 5% for three categories and by 6% for four or five categories. In sum, the dichotomized proxy confounder, representing a dichotomous underlying state, minimized the type-I error rate and performed worst when left as continuous.

Online interactive results tool

For a further analysis of our results, we propose an interactive online application that allows users to manipulate the different parameters used in this study to assess their impact on the Type-I error rate, represented graphically. The application can be accessed through: https://usherbrookeprimus.shinyapps.io/resultsApp/.

Discussion

The results of these simulations confirm and expand the general conclusions of other authors: categorizing a continuous confounding variable leads to a surprisingly large and robust inflation of the Type-I error rate, nearly regardless of model parameters. Only with a very weak association between the confounder and either the outcome or the exposure (i.e., in the absence of a real confounding effect) did this inflation disappear; under many realistic scenarios, the Type-I error was 100%. When applied across hundreds or thousands of studies, even a small inflation of the Type-I error rate – from the expected 5% to, say, 10% – should have a large impact on our confidence in the results generated by a body of literature, especially given the many other biases that tend to lead toward false positive results [23]. The Type-I error rates observed here suggest the problem may be much larger than this small inflation, given the pervasiveness of categorization of important confounders such as age, socio-economic status, and many others.

We identified one highly specific case where categorization diminished the Type-I error, and it is a case chosen specifically to be the exception that proves the rule. This case is when the outcome (i.e., dependent) variable is determined not by the measured confounding variable, but by an underlying dichotomous process for which the measured confounder is a proxy. (Real-world examples might be using blood glucose level to determine diabetes status, or identifying a patient's sex, when unknown, using levels of steroid sex hormones, when it is diabetes status or sex rather than glucose level or hormone level that affects the outcome.) Even in this case, categorization only reduced Type-I error rate relative to the continuous proxy, and when the number of categories corresponded to the number of underlying groups (i.e., 2). And even when all these criteria were met, Type-I error was still substantially higher than the expected α=0.05, reaching error rates greater than 50% under some scenarios.

This counter-example is an example of the principle that all measurement error of a confounding variable increases the risk of Type-I error [21]. In the case of the counter-example, the true variable that should have been measured is the underlying (latent) dichotomous variable, and using a continuous proxy introduces measurement error which can be partially but not completely eliminated by dichotomizing the proxy. The conditions for categorization are thus highly restrictive (and thus may never be met in practice) – one would need to know *a priori* (a) that the continuous variable was a proxy for a true categorical variable, (b) exactly how many underlying categories (sometimes referred to as “latent classes”) there were, and (c) that it was the underlying variable rather than the proxy that was the true confounder. Because confounding variables are generally measured with some measurement error to begin with, the effect of the categorization is over and above the Type-I error inflation due to the original measurement error [21].

The details of our results offer some guidance as to which situations present the greatest Type-I error inflation due to categorization. Type-I error inflation is worse when fewer categories are used. Stronger associations between the confounder and either the exposure or the outcome rapidly increase the Type-I error. Counter-intuitively, large sample size also makes the problem worse, increasing the power to detect the residual confounding present when a confounder is imperfectly measured. All of these effects are quite large.

The effects of the confounder’s distribution and the categorization method are more nuanced. When the confounder has a normal or log-normal distribution, the maximized R^2 method performs worse than the standard quantile method. However, maximized R^2 performed substantially better than quantile under the mixture distribution, a special case when two categories also performs better than more categories. This case demonstrates the limits of simulations for inferring the precise error rate in cases where particular but unknown data generating processes are likely to underlie data structure. In theory, it might be possible to use *a priori* clinical knowledge to slightly diminish the Type-I error rate by choosing optimal cut-offs based on (a) the relationship between the confounder and the outcome; (b) the relationship between the confounder and the exposure, and (c) knowledge of underlying biological/ sociological/ psychological processes. In practice, such *a priori* knowledge is unlikely to be sufficient. Our mathematical derivation of the estimator bias shows substantial complexity in the interactions between such factors and therefore how difficult the task of theoretically controlling for the introduction of measurement error becomes. Traditional clinical cut-offs are unlikely to be valid, for example, unless they approximate underlying biological thresholds, or unless there are threshold effects in their relationships with the other variables. Also, we note that even the best-case scenario for such dichotomization in our simulations still produced substantial Type-I error; such error is unavoidable under the mixture distribution, where the true confounder is unmeasured and an imperfect proxy is used. Even the use of a raw continuous confounding variable in a regression model may sometimes be insufficient: if the relationship of the confounder with the outcome is non-linear, there may still be substantial residual confounding [24]. Quadratic regression, fractional polynomials [25], non -parametric regression [26], and splines are potential solutions to this problem.

All of which is to say that categorization is, in general, a conscious and unnecessary introduction of measurement error. In lay terms, to drive the point home, categorizing a continuous confounder is the equivalent of saying, “Hey, my study is pretty good, but what it could really use is some measurement error. Why don’t I categorize the confounders? That way I will be essentially assured of detecting a positive result whether or not one exists!" In order to help researchers understand the magnitude of the problem, we propose a second interactive online application that allows the users, manually or with the use of the quantile categorization method, to choose cut-offs and assess the probable Type-I error rate of an unrelated exposure controlled for the given categorized confounder. The user can also choose between the distributions and the models presented in this study. The application can be accessed through: https://usherbrookeprimus.shinyapps.io/simulationApp/. Our hope is that this tool will allow many researchers to simulate a situation similar enough to their research question that they get a sense of how bad the problem is likely to be.

Acknowledgments

We thank Shengrui Wang for advice. This research was supported by CIHR grant #s 110789, 120305, 119485 and by NSERC Discovery Grant # 402079-2011. AAC is a member of the FRQ-S-supported *Centre de recherche sur le vieillissement* and *Centre de recherche Clinique du CHUS* and is a funded Research Scholar of the FRQ-S.

References

1. Altman DG, and Royston P. The cost of dichotomising continuous variables. *British Medical Journal* **2006**; **332**, pp. 1080, DOI: 10.1136/bmj.332.7549.1080.

2. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology* **1995**; **6(4)**, pp. 356–365, DOI: 10.1097/00001648-199507000-00005.

3. Morabia A, Ten Have T. Controlling for Continuous Confounders with Non-Gaussian Distribution in Epidemiologic Research. *Epidemiology* **2000**; **11(1)**, pp. 93–94, DOI: 10.2307/3703664.

4. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. *Epidemiology* **1995**; **6(4)**, pp. 450–454, DOI: 10.1097/00001648-199507000-00025.

5. Bennette C, Vickers AJ. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *BMC Medical Research Methodology* **2012**; **12(1)**, pp. 21–25, DOI: 10.1186/1471-2288-12-21.

6. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medecine* **2006**; **25**, pp. 127–141, DOI: 10.1002/sim.2331.

7. Austin PC, Brunner J. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Statistics in Medecine* **2004**; **23**, pp. 1159–1178, DOI: 10.1002/sim.1687.

8. Altman DG. Categorising continuous variables. *British Journal of Cancer* **1991**; **64**, p. 975, DOI: 10.1038/bjc.1991.441.

9. Sigurdsson H, Baldetorp B, Borg A, Dalberg M, Ferno M, Killander D, Olsson H, Ranstam J. Flow cytometry in primary breast cancer: improving the prognostic value of the fraction of cells in the S-phase by optimal categorisation of cut-off levels. *British Journal of Cancer* **1990B**; **62**, pp. 786–790.

10. Taylor JMG, Yu M. Bias and efficiency loss due to categorising an explanatory variable. *Journal of Multivariate Analysis* **2002**; **83**, pp. 248–263, DOI: 10.1006/jmva.2001.2045.

11. Schellingerhout JM, Heymans MW, de Vet HCW, Koes BW, Verhagen AP. Categorizing continuous variables resulted in different predictors in a prognostic model for nonspecific neck pain. *Journal of Clininical Epidemiology* **2009**; **62(8)**, pp. 868–874, DOI: 10.1016/j.jclinepi.2008.10.010.

12. Lausen B, Schumacher M. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Computational Statistics and Data Analysis* **1996**; **21(3)**, pp. 307–326, DOI: 10.1016/0167-9473(95)00016-X.

13. Froslie K, Roislien J, Laake P, Henriksen T, Qvigstad E, Veierod M. Categorisation of continuous exposure variables revisited. A response to the Hyperglycaemia and Adverse Pregnancy Outcome (HAPO) Study. *BMC Medical Research Methodology* **2010**; **10(1)**, p. 103, DOI: 10.1186/1471-2288-10-103.

14. Buettner P, Garbe C, Guggenmoos-Holzmann I. Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma. *Journal of Clininical Epidemiology* **1997**; **50(11)**, pp. 1201–1210, DOI: 10.1016/S0895-4356(97)00155-8.

15. Filardo G, Hamilton C, Hamman B, Ng HKT, Grayburn P. Categorizing BMI may lead to biased results in studies investigating in-hospital mortality after isolated CABG. *Journal of Clininical Epidemiology* **2007**; **60(11)**, pp. 1132–1139, DOI: 10.1016/j.jclinepi.2007.01.008.

16. Groenwold RHH, Klungel OH, Altman DG, van der Graaf Y, Hoes AW, Moons KGM. Adjustment for continuous confounders: an example of how to prevent residual confounding. *Canadian Medical Association Journal* **2013**; **185(5)**, pp. 401–406, DOI: 10.1503/cmaj.120592.

17. Brenner H, Blettner M. Controlling for Continuous Confounders in Epidemiologic Research. *Epidemiology* **1997**; **8(4)**, pp. 429–434, DOI: 10.1097/00001648-199707000-00014.

18. Turner E, Dobson J, Pocock S. Categorisation of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiologic Perspectives and Innovation* **2010**; **7**, p. 9, DOI: 10.1186/1742-5573-7-9.

19. Rothman KJ, Greenland S, Lash TL, *Modern Epidemiology* (3rd edn). Lippincott, Williams & Wilkins: Philadelphia, 2008.

20. Irwin J, McClelland G. Negative consequences of dichotomizing continuous predictor variables. *Journal of Marketing Research* **2003**; **40(3)**, pp. 366-371, DOI: 10.1509/jmkr.40.3.366.19237.

21. Brunner J, Austin PC. Inflation of Type I error rate in multiple regression when independent variables are measured with error. *The Canadian Journal of Statistics* **2009**; **37(1)**, pp. 33-46, DOI: 10.1002/cjs.10004.

22. Li Q. Identifiability of mean-reverting measurement error with instrumental variable. *Statistica Neerlandica* **2014**; **68(2)**, pp. 118–129, DOI: 10.1111/stan.12025.

23. Ioannidis JPA. Why most published research findings are false. *PLoS Medecine* **2005**; **2(8)**, p. e124, DOI: 10.1371/journal.pmed.0020124.

24. Benedetti A, Abrahamowicz M. Using generalized additive models to reduce residual confounding. *Statistics in Medecine* **2004**; **23**, pp. 3781–3801, DOI: 10.1002/sim.2073.

25. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *International Journal of Epidemiology* **1999**; **28(5)**, pp. 964–974, DOI: 10.1093/ije/28.5.964.

26. Rosenberg PS, Katki H, Swanson CA, Brown LM, Wacholder S, Hoover RN. Quantifying epidemiologic risk factors using non-parametric regression: model selection remains the greatest challenge. *Statistics in Medecine* **2003**; **22**, pp. 3369-3381, DOI: 10.1002/sim.1638.

Table 1: Effects of model parameters on Type-1 error rate, modeled separately for (a) confounders with normal, log-normal or bimodal continuous underlying distributions, or (b) confounders with the dichotomous underlying distribution

	Beta	Std. Error	t-value	*p*
Normal, log-normal or bimodal confounder				
Intercept	-0.67	0.018	-38.1	<0.0001
# of categories (numeric variable)	-0.08	0.004	-21.4	<0.0001
Confounder-exposure correlation	0.56	0.008	67.1	<0.0001
Regression Type				
Logistic (ref)	0	-	-	-
Linear	0.09	0.005	19.3	<0.0001
Confounder-outcome association	0.09	0.002	52.0	<0.0001
Sample size/100[a]	0.10	0.002	49.1	<0.0001
Confounder distribution				
Normal (ref)	0	-	-	-
Log-normal	0.003	0.016	0.21	0.84
Bimodal	-0.16	0.016	-10.0	<0.0001
Categorization method				
Quantile (ref)	0	-	-	-
Max R^2	-0.009	0.008	-1.0	0.30
Interaction: # Cat*Distribution[b]				
Log-normal	0.009	0.005	1.7	0.10
Bimodal	0.01	0.005	1.9	0.06
Interaction: Distribution*Cat method[c]				
Log-normal* Max R^2	0.03	0.012	2.2	0.03
Bimodal* Max R^2	0.18	0.012	15.2	<0.0001
Dichotomous confounder				
Intercept	-0.86	0.023	-37.5	<0.0001
# of categories				
2 categories (ref)	0	-	-	-
3 categories	0.05	0.014	3.2	0.002
4 categories	0.06	0.014	4.3	<0.0001
5 categories	0.07	0.014	4.6	<0.0001
Confounder-exposure correlation	0.44	0.013	35.2	<0.0001
Regression Type				
Logistic (ref)	0	-	-	-
Linear	0.05	0.007	6.7	<0.0001
Confounder-outcome association	0.10	0.003	38.5	<0.0001
Sample size/100[a]	0.09	0.003	30.6	<0.0001
Categorization method				
Quantile (ref)	0	-	-	-

Max R^2	-0.10	0.014	-7.02	<0.0001
Interaction: # Cat*Cat method= Max R^2				
2 categories (ref)	0	-	-	-
3 categories	0.06	0.020	2.8	0.006
4 categories	0.07	0.020	3.3	0.0009
5 categories	0.07	0.020	3.4	0.0006

[a] This is the effect of the natural logarithm of the continuous sample size on the Type-I error rate.
[b] This is the increase in Type-I error rate per additional category under a log-normal and bimodal distribution.
[c] This is the increase in Type-I error rate with the max R^2 method under a log-normal and bimodal distribution.

Figure legends

Figure 1. Thresholds/cut-offs found by the quantiles (A) and maximized R^2 (B) methods for 2 categories in a sample size of 2000 under the bimodal (3) distribution.

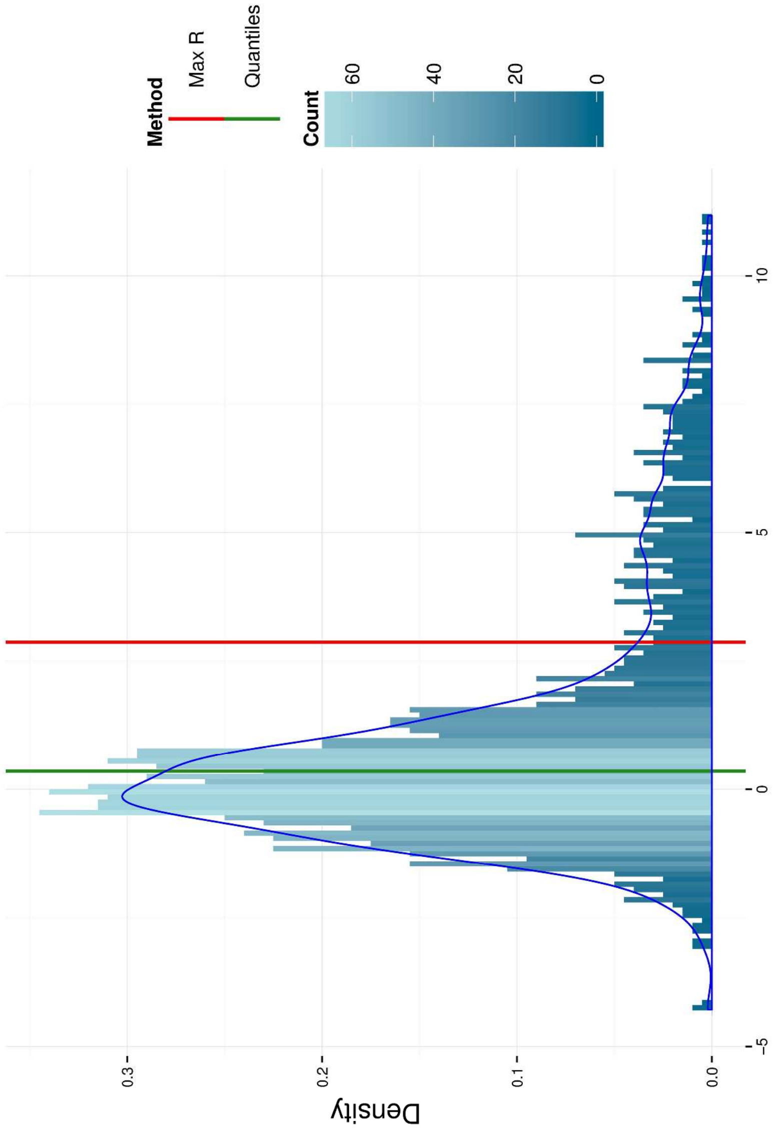
Figure 2. Type-I error rates for logistic (a) models with the confounding variable continuous and in 2-5 categories, using the quantiles (A) method, under normal (1), log-normal (2) and bimodal (3) underlying distributions, $\beta_2 = 2$ and sample size=2000. Vertical lines represent 95% confidence intervals for the simulated Type-1 error rates based on an N of 1000 simulations.

Figure 3. Type-I error rates for linear (b) models with $\sigma^2 = \{0.48, 1.02, 1.73, 3.17, 9.95\}$ and sample size={100, 500, 2000}, using the maximized R^2 (B) method for the confounding variable in 2 categories. Vertical lines represent 95% confidence intervals for the simulated Type-1 error rates based on an N of 1000 simulations.

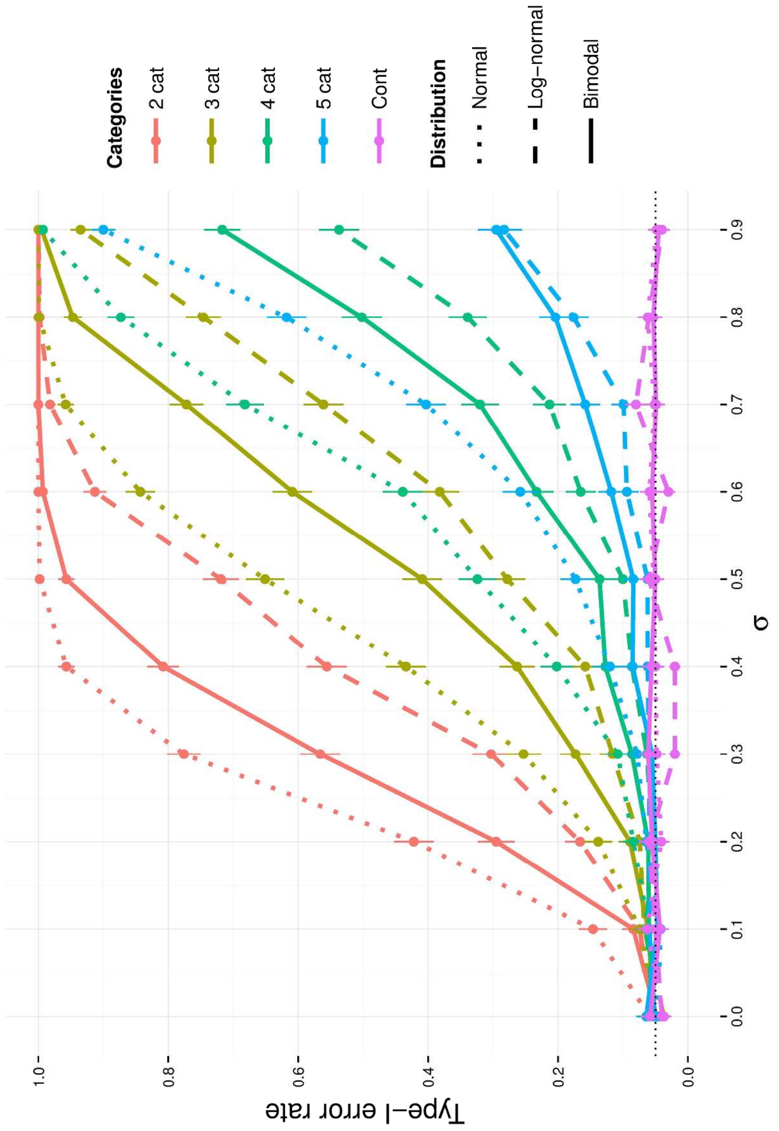
Figure 4. Type-I error rates for logistic (b) models with the confounding variable in 3 categories using the quantiles (A) and the maximized R^2 (B) methods under normal (1), log-normal (2) and bimodal (3) underlying distributions, $\beta_2 = 2$ and sample size=2000.

Vertical lines represent 95% confidence intervals for the simulated Type-1 error rates based on an N of 1000 simulations.

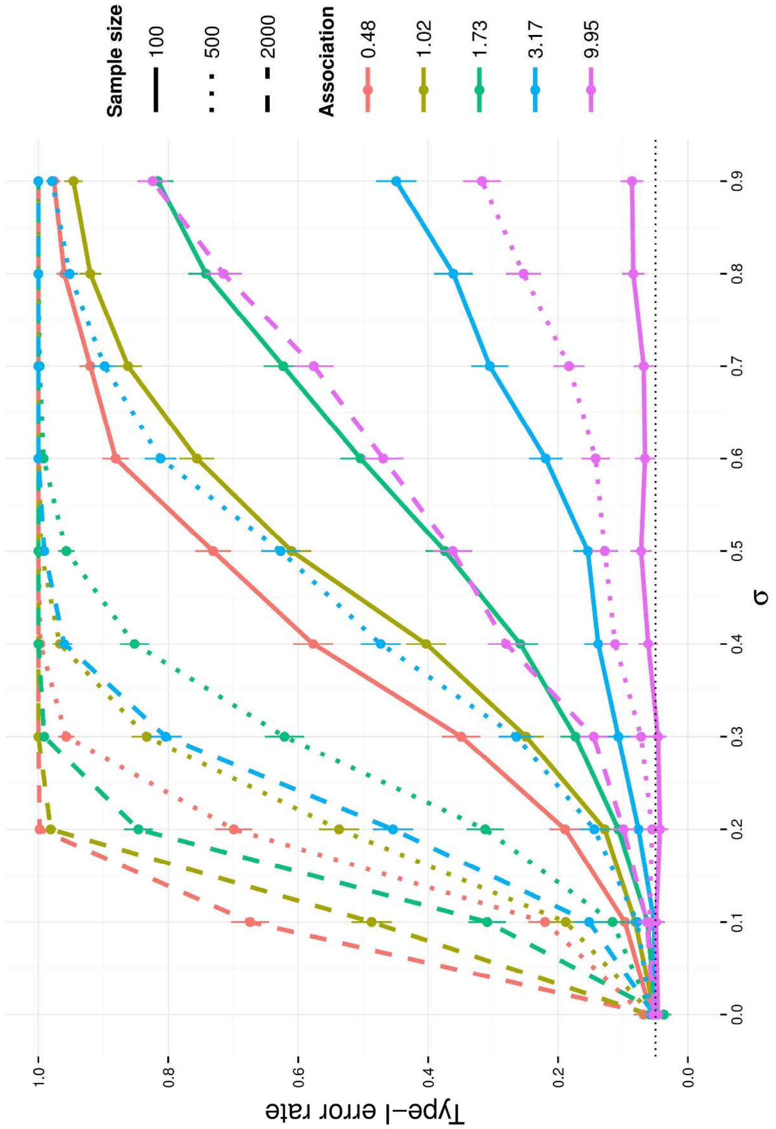
Figure 5. Type-I error rates for linear (b) models with the proxy variable continuous and in 2-5 categories, using the quantiles (A) and the maximized R^2 (B) methods, under dichotomous (4) underlying distribution, $\sigma^2 = 1.02$ and sample size=2000. Vertical lines represent 95% confidence intervals for the simulated Type-1 error rates based on an N of 1000 simulations.



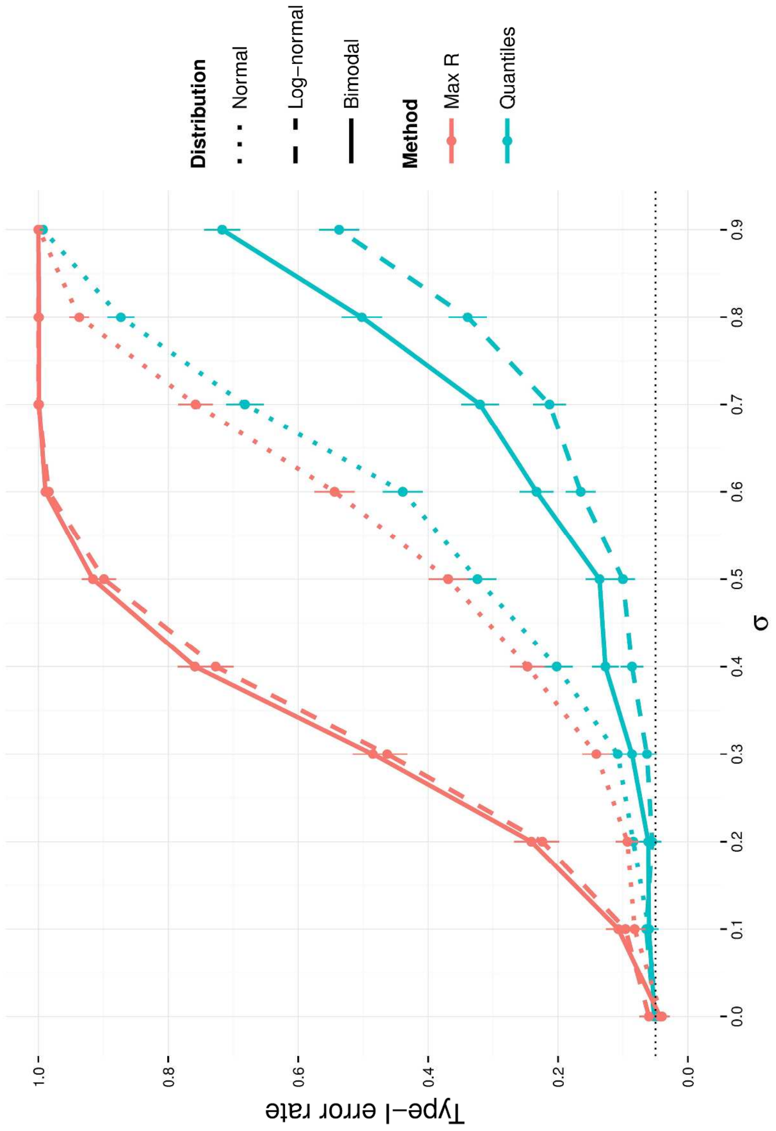
279x361mm (300 x 300 DPI)



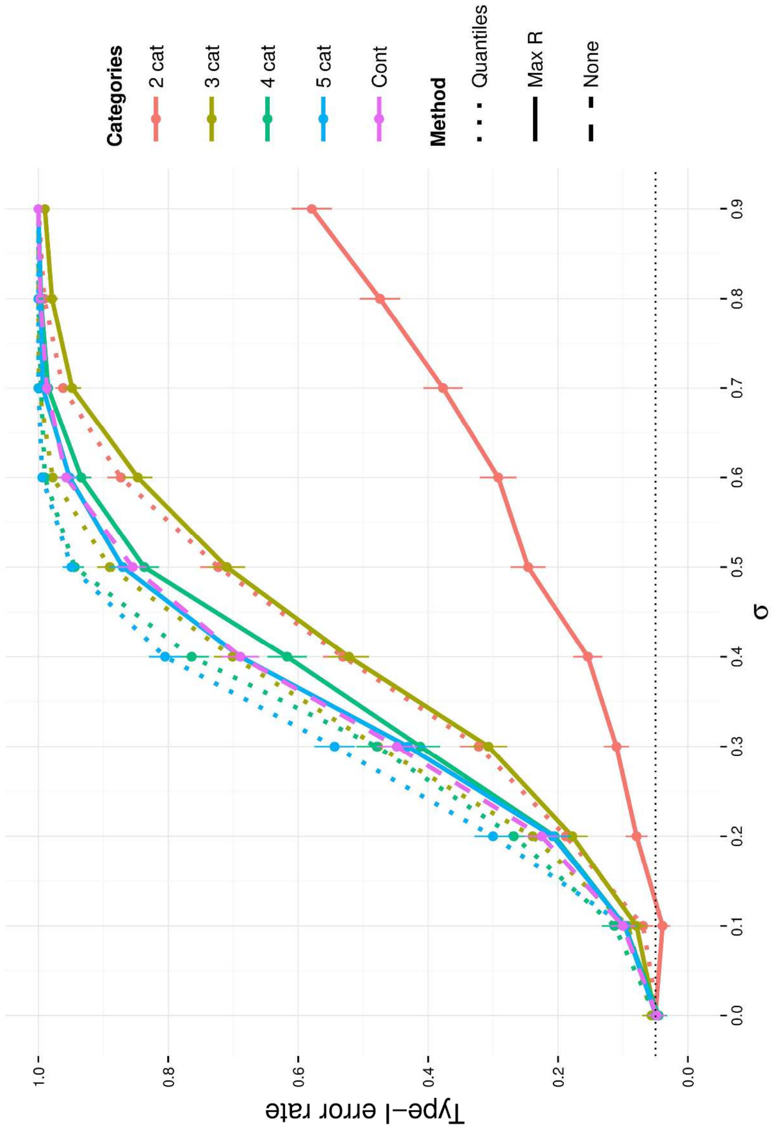
279x361mm (300 x 300 DPI)



279x361mm (300 x 300 DPI)



279x361mm (300 x 300 DPI)



279x361mm (300 x 300 DPI)