Journal of Multivariate Analysis **83**, 248–263 (2002)
doi:10.1006/jmva.2001.2045

Bias and Efficiency Loss Due to Categorizing an Explanatory Variable

Jeremy M. G. Taylor and Menggang Yu

University of Michigan

Received July 25, 2000; published online February 12, 2002

It is a common situation in biomedical research that one or more variables are known to be associated with the outcome of interest. Researchers often discretize some variables and fit a regression model using these discretized variables. Although convenient for illustration purposes, such an approach can be biased and lead to loss of efficiency. In this article, we consider the situation of a regression model with two explanatory variables under an assumption of multivariate normality. We investigate the effect of dichotomizing or categorizing one variable on the estimate of the coefficient of the other continuous variable and on prediction from the models. Algebraic expressions are presented for the asymptotic bias and variance of the coefficient of the continuous explanatory variable and for the residual sum of squares for prediction. Some numerical examples are presented in which we find that the bias of the coefficient of the continuous explanatory variable is always smaller for the categorized model than that for the dichotomized model. The size of the test of a zero coefficient for the continuous variable only depends on the correlations between the response variable, the discretized variable, and the continuous variable. The size of the test for the categorized model is always smaller than for the dichotomized model, however, both can differ substantially from the nominal level if the correlation between the response and the categorical variable or between the two explanatory variables is high. The (predictive) relative efficiency of models also only depends on correlations amongst the three variables. There is a substantial loss of efficiency due to categorization if the correlation between the categorized and response variable is high. The predictive relative efficiency is always higher for the categorized model. The relative predictive efficiency due to dichotomization depends on the choice of cut points, with the least loss of efficency being achieved at the median.

AMS 2000 subject classifications: 62J05; 62F05; 62F12; 62H12.

Key words and phrases: cutpoints; discretization; regression.

1. INTRODUCTION

Biomedical researchers are often interested in assessing the relationship or association between a response variable and a number of explanatory variables. This is frequently achieved by building a statistical regression model and then potentially using this model for prediction. General

0047-259X/02 $35.00

approaches to model building are discussed in Harrell *et al.* (1996). A common occurrence is for the investigator to categorize a continuous variable before the model is developed. This is done because there may be uncertainty about the appropriate functional form for the continuous variable and the belief that discretizing it will impart robustness on the conclusion of the analysis. Furthermore it is generally easier to display and explain results with discrete, rather than continuous variables. Many authors, particularly in areas of application, seem to believe the desirability of dichotomizing continuous variables. Apart from the throwing away of information, which is the topic of the current paper, this procedure produces a possibly scientifically unrealistic model where the effect has a sudden jump at the cut-off value, with all values below the cut-off having equal effect and all values below the cutoff having equal effect (Altman, 1991).

How a continuous variable is discretized and some of the implications of discretization have been considered by a number of authors. Cox (1957) gives a general discussion of grouping with the goal of retaining as much information as possible. For the case of a single explanatory variable Lagakos (1988) studied the asymptotic relative efficiency of tests of zero regression coefficient, comparing the continuous (Z) versus the discrete (Z_x) variables. He showed that the loss of efficiency is given by the square of the correlation between Z and Z_x. Connor (1972) considered the choice of optimal cut points for a continuous variable. Zhao and Kolonel (1992) considered in a simulation study the efficiency loss in the odds ratio due to categorizing a continuous variable in a case-control study. Morgan and Elashoff (1986) considered the effect of categorizing a continuous covariate on the estimated hazard ratio between two groups in survival analysis. They showed how the loss of efficiency depends on the number of categories and the choice of cutpoints. Altman *et al.* (1994) and Lausen and Schumacher (1996) considered choosing optimal cutpoints and demonstrated the dangers of inference following the data-driven selection of cutpoints.

In this paper we consider the situation of two explanatory variables in a regression model. It is a common situation in biomedical research that one or more variables are known to be associated with the outcome. There are also other variables, perhaps coming from a newly developed technique or assay, which are thought also to be associated with the response variable. For example, in prostate cancer PSA is known to be associated with survival after therapy. An important question in such a setting is does a new variable add any information about survival over and above what is contained in the first variable (i.e., PSA). A common approach to this problem is to discretize the first variable and then fit a regression model with the discretized first variable and the new variable as the two explanatory variables. A significant coefficient for the new variable would typically be

interpreted as indicating that this variable had some independent prognostic importance. In this paper we address whether this may be due to the discretization of the first variable. A second related issue concerns prediction; it may be that there is negligible loss of efficiency associated with prediction using a discretized version of the first variable plus the new variable compared to using a continuous version of the first variable. The new variables could be, in effect, recovering the information lost due to the discretization of the first variable.

The algebraic development in this paper is based on the multivariate normal model. Section 2 contains results concerning bias and variance of parameter estimates, Section 3 contains results concerning prediction and Section 4 contains results concerning the size of tests. Section 5 contains numerical examples.

2. BIAS AND VARIANCE OF PARAMETER ESTIMATES

Assume $(Y, X_1, X_2)'$ has a multivariate normal distribution with mean vector $(\mu_y, \mu_{x_1}, \mu_{x_2})'$ and covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{yx_1} & \sigma_{yx_2} \\ \sigma_{yx_1} & \sigma_{x_1}^2 & \sigma_{x_1x_2} \\ \sigma_{yx_2} & \sigma_{x_1x_2} & \sigma_{x_2}^2 \end{pmatrix}.$$

We regard Y as the response variable, X_1 as the variable which is known to be associated with Y, and X_2 as the new variable which is assumed independent of Y given X_1. Thus X_2 is a variable which is not important if X_1 is known, but may appear to be important if X_1 is discretized. A consequence of the assumption $[Y \mid x_1, x_2] \equiv [Y \mid x_1]$, is that

$$\sigma_{yx_1}\sigma_{x_1x_2} = \sigma_{yx_2}\sigma_{x_1}^2$$

or equivalently, in terms of correlation,

$$\rho_{yx_1}\rho_{x_1x_2} = \rho_{yx_2}. \tag{1}$$

The true regression model for the response Y given X_1 and X_2 is

$$Y = \alpha_0 + \alpha_1 X_1 + e, \tag{2}$$

where $e \sim N(0, \sigma_y^2 - \sigma_{yx_1}^2/\sigma_{x_1}^2)$, $\alpha_1 = \sigma_{yx_1}/\sigma_{x_1}^2 = \rho_{yx_1}\sigma_y/\sigma_{x_1}$, and $\alpha_0 = \mu_y - \alpha_1\mu_{x_1}$. Without loss of generality, we assume $\mu_y = \mu_{x_1} = \mu_{x_2} = 0$.

2.1. *The Dichotomized Model*

First we consider the simplest case where X_1 is dichotomized at the median. Assume we fit the model of the form $Y = \beta_0^D + \beta_1^D X_1^* + \beta_2^D X_2 + e$, where X_1^* is X_1 dichotomized at point 0,

$$X_1^* = \begin{cases} 1 & \text{if } X_1 \geqslant 0; \\ -1 & \text{if } X_1 < 0. \end{cases}$$

It is easy to show that

$$\mathrm{Var}(X_1^*) = 1, \qquad \mathrm{cov}(X_1^*, X_2) = \sqrt{\frac{2}{\pi}}\,\frac{\sigma_{x_1x_2}}{\sigma_{x_1}},$$

$$\mathrm{cov}(X_1^*, Y) = \sqrt{\frac{2}{\pi}}\,\frac{\sigma_{x_1y}}{\sigma_{x_1}}, \qquad \text{and} \qquad \mathrm{cov}(X_1^*, X_1) = \sqrt{\frac{2}{\pi}}\,\sigma_{x_1}.$$

Let $\mathbf{y}$, $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_1^*$ denote the set of n observations. Then the least-square estimate of β_2^D is

$$\widehat{\beta_2^D} = \frac{s_{\mathbf{x}_1^*}^2 \widehat{\mathrm{cov}}(\mathbf{x}_2, \mathbf{y}) - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)\,\widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{y})}{s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_2}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)^2},$$

where $s_{\mathbf{x}_1^*}^2 = \frac{1}{n}\sum(x_{1i}^* - \overline{x_1^*})^2$ and $\widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2) = \frac{1}{n}\sum(x_{1i}^* - \overline{x_1^*})(x_{2i} - \bar{x}_2)$, and other covariances are similarly defined.

To find expressions for the mean and variance of $\widehat{\beta_2^D}$ we adapt the approach given in Lagakos (1988). The details of the derivations are given in the Appendix.

Define $U(\mathbf{x}_1, \mathbf{x}_2)$ by

$$\begin{aligned} U(\mathbf{x}_1, \mathbf{x}_2) &\equiv E(\widehat{\beta_2^D} \mid \mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2) \\ &= \frac{s_{\mathbf{x}_1^*}^2 \widehat{\mathrm{cov}}(\mathbf{x}_1, \mathbf{x}_2) - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)\,\widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_1)}{s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_2}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)^2}\,\frac{\sigma_{x_1y}}{\sigma_{x_1}^2}. \end{aligned} \tag{3}$$

For large n, $U(\mathbf{x}_1, \mathbf{x}_2)$ converges to

$$\frac{1 \cdot \sigma_{\mathbf{x}_1\mathbf{x}_2} - \sqrt{\frac{2}{\pi}}\,\frac{\sigma_{x_1x_2}}{\sigma_{x_1}} \cdot \sqrt{\frac{2}{\pi}}\,\sigma_{x_1}}{\sigma_{x_2}^2 - \left(\sqrt{\frac{2}{\pi}}\,\frac{\sigma_{x_1x_2}}{\sigma_{x_1}}\right)^2}\,\frac{\sigma_{x_1y}}{\sigma_{x_1}^2} = \frac{0.3634}{1 - 0.6366\rho_{x_1x_2}^2}\,\frac{\sigma_{x_2y}}{\sigma_{x_2}^2}$$

Thus the unconditional expectation of $\widehat{\beta_2^D}$ is

$$E(\widehat{\beta_2^D}) = \frac{0.3634}{1-0.6366\rho_{x_1x_2}^2}\frac{\sigma_{x_2y}}{\sigma_{x_2}^2} = \frac{0.3634\rho_{x_2y}}{1-0.6366\rho_{x_1x_2}^2}\frac{\sigma_y}{\sigma_{x_2}}.$$

The conditional variance of $\widehat{\beta_2^D}$ is given by

$$\begin{aligned} V(\mathbf{x_1}, \mathbf{x_2}) &\equiv \mathrm{Var}(\widehat{\beta_2^D} \mid \mathbf{X_1} = \mathbf{x_1}, \mathbf{X_2} = \mathbf{x_2}) \\ &= \frac{\sum_i \left[\frac{1}{n} s_{\mathbf{x}_1^*}^2 (x_{2i} - \overline{\mathbf{x}_2}) - \frac{1}{n}\widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)(x_{1i}^* - \overline{\mathbf{x}_1^*})\right]^2}{[s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_2}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)^2]^2}\sigma_{y\mid x_1, x_2}^2 \\ &= \frac{\frac{1}{n}s_{\mathbf{x}_1^*}^2}{s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_2}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)^2}\sigma_{y\mid x_1}^2. \end{aligned}$$

The unconditional variance of $\widehat{\beta_2^D}$ is $\mathrm{Var}[U(\mathbf{X_1}, \mathbf{X_2})] + E[V(\mathbf{X_1}, \mathbf{X_2})]$, which can be expressed as

$$\begin{aligned} \mathrm{Var}(\widehat{\beta_2^D}) &= E[\mathrm{Var}(\widehat{\beta_2^D} \mid \mathbf{X_1}, \mathbf{X_2})] + \mathrm{Var}[E(\widehat{\beta_2^D} \mid \mathbf{X_1}, \mathbf{X_2})] \\ &= \frac{1-\rho_{x_1y}^2}{1-0.6366\rho_{x_1x_2}^2}\cdot\frac{\sigma_y^2}{n\sigma_{x_2}^2} + \frac{0.3634\rho_{x_1y}^2\sigma_y^2}{n\sigma_{x_2}^2} \end{aligned}$$

2.2. *The Categorized Model*

Now consider the model where X_1 is discretized into three groups of equal probabilities. Assume we fit the model $Y = \beta_0^C + \beta_{11}^C X_{11}^* + \beta_{12}^C X_{12}^* + \beta_2^C X_2 + e$, where

$$X_{11}^* = \begin{cases} \frac{2}{3} & \text{if } X_1 \leqslant -d; \\ -\frac{1}{3} & \text{if } X_1 > -d. \end{cases}$$

$$X_{12}^* = \begin{cases} \frac{2}{3} & \text{if } X_1 \geqslant d; \\ -\frac{1}{3} & \text{if } X_1 < d. \end{cases}$$

Then we have the following equalities:

$$\mathrm{Var}(X_{11}^*) = \mathrm{Var}(X_{12}^*) = \frac{2}{9}, \mathrm{cov}(X_{11}^*, X_{12}^*) = -\frac{1}{9},$$

$$\mathrm{cov}(X_{11}^*, X_2) = -\mathrm{cov}(X_{12}^*, X_2) = -0.3637\frac{\sigma_{x_1x_2}}{\sigma_{x_1}},$$

$$\mathrm{cov}(X_{11}^*, X_1) = -\mathrm{cov}(X_{12}^*, X_1) = -0.3637\sigma_{x_1}.$$

The least-square estimate of β_2^C is,

$$\widehat{\beta_2^C} = \frac{U\,\widehat{\mathrm{cov}}(\mathbf{x}_{11}^*, \mathbf{y}) + V\,\widehat{\mathrm{cov}}(\mathbf{x}_{12}^*, \mathbf{y}) + W\,\widehat{\mathrm{cov}}(\mathbf{x}_2, \mathbf{y})}{U\,\widehat{\mathrm{cov}}(\mathbf{x}_{11}^*, \mathbf{x}_2) + V\,\widehat{\mathrm{cov}}(\mathbf{x}_{12}^*, \mathbf{x}_2) + W s_{\mathbf{x}_2}^2}$$

where

$$U = \widehat{\mathrm{cov}}(\mathbf{x}_{11}^*, \mathbf{x}_{12}^*)\,\widehat{\mathrm{cov}}(\mathbf{x}_{12}^*, \mathbf{x}_2) - s_{\mathbf{x}_{12}^*}^2\,\widehat{\mathrm{cov}}(\mathbf{x}_{11}^*, \mathbf{x}_2) \to 0.0404\,\frac{\sigma_{x_1x_2}}{\sigma_{x_1}},$$

$$V = \widehat{\mathrm{cov}}(\mathbf{x}_{11}^*, \mathbf{x}_{12}^*)\,\widehat{\mathrm{cov}}(\mathbf{x}_{11}^*, \mathbf{x}_2) - s_{\mathbf{x}_{11}^*}^2\,\widehat{\mathrm{cov}}(\mathbf{x}_{12}^*, \mathbf{x}_2) \to -0.0404\,\frac{\sigma_{x_1x_2}}{\sigma_{x_1}},$$

$$W = s_{\mathbf{x}_{11}^*}^2 s_{\mathbf{x}_{12}^*}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_{11}^*, \mathbf{x}_{12}^*)^2 \to \frac{1}{27}.$$

Using similar algebra to that in Section 2.1 we have for large n that

$$E(\widehat{\beta_2^C}) \approx \frac{0.206}{1-0.794\rho_{x_1x_2}^2} \cdot \frac{\sigma_{x_2y}}{\sigma_{x_2}^2} = \frac{0.206\rho_{x_2y}}{1-0.794\rho_{x_1x_2}^2}\,\frac{\sigma_y}{\sigma_{x_2}},$$

and

$$\mathrm{Var}(\widehat{\beta_2^C}) = \frac{1-\rho_{x_1y}^2}{1-0.794\rho_{x_1x_2}^2} \cdot \frac{\sigma_y^2}{n\sigma_{x_2}^2} + 0.206\rho_{x_1y}^2 \cdot \frac{\sigma_y^2}{n\sigma_{x_2}^2}.$$

2.3. *The Unbalanced Dichotomized Model*

We now consider the situation where X_1 is dichotomized at an arbitrary point d, not necessarily the median. Let

$$X_1^* = \begin{cases} \Phi\left(\dfrac{d}{\sigma_{x_1}}\right) & \text{if} \quad X_1 \geqslant d \\ -\left(1-\Phi\left(\dfrac{d}{\sigma_{x_1}}\right)\right) & \text{if} \quad X_1 < d, \end{cases}$$

where Φ is the Gaussian distribution function. Then we have

$$\mathrm{cov}(X_1^*, X_2) = 2\phi\left(\frac{d}{\sigma_{x_1}}\right)\frac{\sigma_{x_1x_2}}{\sigma_{x_1}}$$

$$\mathrm{cov}(X_1^*, Y) = 2\phi\left(\frac{d}{\sigma_{x_1}}\right)\frac{\sigma_{x_1y}}{\sigma_{x_1}}$$

$$\mathrm{cov}(X_1^*, X_1) = 2\phi\left(\frac{d}{\sigma_{x_1}}\right)\sigma_{x_1}$$

$$\mathrm{Var}(X_1^*) = 4pq,$$

where $p = \Phi(d/\sigma_{x_1})$, $q = 1-p$, and ϕ is the Gaussian density function.

Fitting a model of the form $Y = \beta_0^D + \beta_1^D X_1^* + \beta_2^D X_2 + e$, we can show that for large n,

$$E(\hat{\beta}_2^D) = \frac{pq - \phi^2\left(\frac{d}{\sigma_{x_1}}\right)}{pq - \phi^2\left(\frac{d}{\sigma_{x_1}}\right)\rho_{x_1x_2}^2} \frac{\rho_{x_2y}\sigma_y}{\sigma_{x_2}}$$

$$\mathrm{Var}(\widehat{\beta_2^D}) = \frac{1 - \frac{\phi^2\left(\frac{d}{\sigma_{x_1}}\right)}{pq}}{4pq} \cdot \frac{\rho_{x_1y}^2\sigma_y^2}{n\sigma_{x_2}^2} + \frac{1-\rho_{x_1y}^2}{1 - \frac{\phi^2\left(\frac{d}{\sigma_{x_1}}\right)}{pq}\rho_{x_1x_2}^2} \cdot \frac{\sigma_y^2}{n\sigma_{x_2}^2}$$

An obvious feature of all these categorized models is that $E(\beta_2^D) \neq 0$. Furthermore, the bias increases as both the correlation between X_2 and Y and the correlation between X_1 and X_2 increases.

3. EFFICIENCY OF PREDICTIONS

Even though categorizing a continuous variable leads to biased parameter estimates, it is conceivable that there is negligible loss of efficiency in predictions. In this paper we use the residual sum of squares as a measure of efficiency. We compare residual sums of squares of different models with the residual sum of squares from the continuous model $Y = \alpha_0 + \alpha_1 X_1 + e$.

In vector notation, let $\mathbf{r} = \mathbf{Y} - \hat{\beta}\mathbf{X}$ denote the residual for a general regression model. A well known property of the residual sum of squares (Graybill, 1976) is that

$$(\sigma_{y|x_1}^2)^{-1}\,\mathbf{r}'\mathbf{r} = (\sigma_{y|x_1}^2)^{-1}\,\mathbf{Y}'(\mathbf{I} - \mathbf{X}\mathbf{X}^-)\,\mathbf{Y}$$

has a noncentral chi-square distribution with d.f. $= n-p$ and noncentrality $\lambda = \frac{1}{2}\mu_{y|x}'(\mathbf{I} - \mathbf{X}\mathbf{X}^-)\,\mu_{y|x}$, where, $\sigma_{y|x_1}^2 = \sigma_y^2(1-\rho_{x_1y}^2)$ and $\mu_{y|x_1} = \mu_y + \alpha_1(x_1 - \mu_{x_1})$ is the conditional variance and mean of Y given X_1.

Applying this result we have for the continuous model $(\sigma_{y|x_1}^2)^{-1}\,\mathbf{r}'\mathbf{r} \sim \chi_{n-2}^2$.

For the dichotomized model, the noncentrality parameter is

$$\lambda = \frac{\sigma_{x_1y}^2}{2\sigma_{x_1}^4} \cdot \left[\mathbf{X}_1'\mathbf{J}_n\mathbf{X}_1 - \frac{\begin{pmatrix} n\hat{s}_{x_2}^2\,\widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_1)^2 + n\hat{s}_{x_1}^{2*}\,\widehat{\mathrm{cov}}(\mathbf{x}_1, \mathbf{x}_2)^2 \\ -2n\,\widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)\,\widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_1)\,\widehat{\mathrm{cov}}(\mathbf{x}_1, \mathbf{x}_2) \end{pmatrix}}{s_{x_1}^{2*}s_{x_2}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_2)^2} \right],$$

where J_n is the $n \times n$ matrix with all entries equal to 1. For large n, it can be shown that

$$\frac{1}{n}\lambda \to \frac{1}{2}\cdot\frac{\sigma_{x_1 y}^2}{\sigma_{x_1}^2}\left(\frac{(\pi-2)(1-\rho_{x_1x_2}^2)}{\pi-2\rho_{x_1x_2}^2}\right)$$
$$= \frac{1}{2}\cdot\sigma_y^2\rho_{x_1y}^2\left(\frac{0.3634(1-\rho_{x_1x_2}^2)}{1-0.6366\rho_{x_1x_2}^2}\right)$$

Therefore the relative efficiency is given by

$$\frac{E(\mathbf{r}'\mathbf{r}\mid \mathbf{X}_1)}{E(\mathbf{r}'\mathbf{r}\mid \mathbf{X}_1^*, \mathbf{X}_2)} \approx \frac{1}{1+\dfrac{\rho_{x_1y}^2}{1-\rho_{x_1y}^2}\cdot\dfrac{0.3634(1-\rho_{x_1x_2}^2)}{1-0.6366\rho_{x_1x_2}^2}}. \tag{4}$$

Note that the denominator is always greater than or equal to one, indicating loss of efficiency unless X_1 and Y are uncorrelated or X_1 and X_2 are perfectly correlated.

For the three-category model, it can be shown that

$$\frac{1}{n}\lambda \to \frac{1}{2}\cdot\sigma_y^2\rho_{x_1y}^2\cdot\frac{0.206(1-\rho_{x_1x_2}^2)}{1-0.794\rho_{x_1x_2}^2},$$

giving relative efficiency

$$\frac{E(\mathbf{r}'\mathbf{r}\mid \mathbf{X}_1)}{E(\mathbf{r}'\mathbf{r}\mid \mathbf{X}_{11}^*, \mathbf{X}_{12}^*, \mathbf{X}_2)} \approx \frac{1}{1+\dfrac{\rho_{x_1y}^2}{1-\rho_{x_1y}^2}\cdot\dfrac{0.206(1-\rho_{x_1x_2}^2)}{1-0.794\rho_{x_1x_2}^2}}. \tag{5}$$

For the general dichotomized model,

$$\frac{1}{n}\lambda \to \frac{1}{2}\cdot\frac{\sigma_{x_1y}^2}{\sigma_{x_1}^4}\cdot\sigma_{x_1}^2\cdot\frac{\left(1-\dfrac{\phi^2\left(\dfrac{d}{\sigma_{x_1}}\right)}{pq}\right)(1-\rho_{x_1x_2}^2)}{1-\dfrac{\phi^2\left(\dfrac{d}{\sigma_{x_1}}\right)}{pq}\rho_{x_1x_2}^2},$$

giving relative efficiency

$$\frac{E(\mathbf{r}'\mathbf{r} \mid \mathbf{X}_1)}{E(\mathbf{r}'\mathbf{r} \mid \mathbf{X}_1^*, \mathbf{X}_2)} \approx \frac{1}{1 + \dfrac{\dfrac{\rho_{x_1y}^2}{1-\rho_{x_1y}^2} \cdot \left(1 - \dfrac{\phi^2\left(\dfrac{d}{\sigma_{x_1}}\right)}{pq}\right)(1-\rho_{x_1x_2}^2)}{1 - \dfrac{\phi^2\left(\dfrac{d}{\sigma_{x_1}}\right)}{pq}\rho_{x_1x_2}^2}}. \tag{6}$$

In all cases we note that the relative efficiencies depend only on the correlation between X_1 and Y and between X_1 and X_2.

4. SIZE OF THE TEST

A common approach to determining whether X_2 is important would be to test H_0: $\beta_2 = 0$ vs. H_a: $\beta_2 > 0$. For any particular model, the size of a nominal 0.05 level test is determined by

$$\Phi\left(\frac{\hat{\mu}_{\beta_2}}{\sqrt{\hat{\sigma}_{\beta_2}^2}} - 1.65\right),$$

where $\hat{\mu}_{\beta_2}$ and $\hat{\sigma}_{\beta_2}^2$ are the asymptotic mean and variance of β_2 in the corresponding dichotomized or categorized model. For the dichotomized model, the size is

$$\Phi\left(\sqrt{n}\,\frac{\dfrac{0.3634\rho_{x_2y}}{1-0.6366\rho_{x_1x_2}^2}}{\sqrt{\dfrac{1-\rho_{x_1y}^2}{1-0.6366\rho_{x_1x_2}^2} + 0.3634\rho_{x_1y}^2}} - 1.65\right). \tag{7}$$

For the categorized model, the size is

$$\Phi\left(\sqrt{n}\,\frac{\dfrac{0.206\rho_{x_2y}}{1-0.794\rho_{x_1x_2}^2}}{\sqrt{\dfrac{1-\rho_{x_1y}^2}{1-0.794\rho_{x_1x_2}^2} + 0.206\rho_{x_1y}^2}} - 1.65\right). \tag{8}$$

Recalling from (1) that ρ_{x_2y} can be written as $\rho_{x_1y}\rho_{x_1x_2}$, we can see that the size only depends on the correlation between Y and X_1 and the correlation between X_1 and X_2.

5. NUMERICAL RESULTS

5.1. *Size*

We undertook a small simulation study, to compare with the asymptotic results, and to illustrate the effect of discretizing X_1 on the bias and variance of β_2 and on the size of tests of $\beta_2 = 0$. The simulation results were produced using 1,000 replicates and a sample size of 200 with data generated from the true model (2). The design is chosen such that $\sigma^2_{x_1} = \sigma^2_{x_2} = 100$ and $\sigma^2_y = 10,000$.

We fit three models

$$Y = \beta_0^D + \beta_1^D X_1^* + \beta_2^D X_2 + e \tag{9}$$

$$Y = \beta_0^c + \beta_{11}^c X_{11}^* + \beta_{12}^c X_{12}^* + \beta_2^c X_2 + e \tag{10}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e. \tag{11}$$

In model (9), X_1 is dichotomized at the median. In model (10), X_1 is categorized into three groups of equal probability. Model (11) is fitted without dichotomizing X_1. Model (11) is fully efficient for large samples, with

$$E(\hat{\beta}_2) = 0 \qquad \mathrm{Var}(\hat{\beta}_2) = \frac{\sigma^2_y}{n\sigma^2_{x_2}} \frac{1-\rho^2_{x_1 y}}{1-\rho^2_{x_1 x_2}}.$$

Table I shows the mean and variance of $\hat{\beta}_2$ for the three models, for nine different configurations of $\rho_{x_1 x_2}$ and $\rho_{x_1 y}$. We see excellent correspondence between the asymptotic and the simulation results. The bias of $\hat{\beta}_2$ is larger for the dichotomized model than for the categorized model. The comparisons of the variance of $\hat{\beta}_2$ are mixed, with no one model always giving the smallest or the largest variance. They are equal only if the correlation between X_1 and X_2 is 1.

Figure 1 shows the size of the test of $\beta_2 = 0$ as a function of $\rho_{x_1 x_2}$ calculated from Eqs. (7) and (8). We see that the size of the test quickly departs from the nominal 0.05 level for both dichotomized and categorized models. Furthermore the size is uniformly smaller for the categorized model than for the dichotomized model. When the correlation between X_1 and X_2 is high, they both have a similar same large size. This is to be expected since rejecting $\beta_2 = 0$ is not hard since the high correlation between X_2 and X_1 implies the high correlation of X_2 with Y given fixed $\rho_{x_1 y}$ (from (1)). The size departs more quickly from the nominal level when the correlation between X_1 and Y is higher. We also note that the size does not depend on the variances.

TABLE I

Simulation Results for the Mean and Variance of β_2 under Different Models

configuration			Model with X_2(Eq. 11). asymp.	simul.	Dichotomized model (Eq. 9). asymp.	simul.	Categorized model (Eq. 10). asymp.	simul.
		Mean(β_2)	0.000	0.031	0.347	0.381	0.200	0.228
$\rho_{x_1x_2}=0.3$	$\rho_{x_1y}=0.3$	Var(β_2)	0.500	0.507	0.499	0.501	0.499	0.502
		Mean(β_2)	0.000	−0.023	0.578	0.546	0.333	0.310
$\rho_{x_1x_2}=0.3$	$\rho_{x_1y}=0.5$	Var(β_2)	0.412	0.420	0.443	0.461	0.430	0.448
		Mean(β_2)	0.000	0.008	0.925	0.927	0.533	0.542
$\rho_{x_1x_2}=0.3$	$\rho_{x_1y}=0.8$	Var(β_2)	0.198	0.212	0.307	0.311	0.260	0.281
		Mean(β_2)	0.000	−0.002	0.648	0.631	0.386	0.366
$\rho_{x_1x_2}=0.5$	$\rho_{x_1y}=0.3$	Var(β_2)	0.607	0.623	0.557	0.578	0.577	0.596
		Mean(β_2)	0.000	0.020	1.081	1.106	0.643	0.662
$\rho_{x_1x_2}=0.5$	$\rho_{x_1y}=0.5$	Var(β_2)	0.500	0.521	0.491	0.513	0.494	0.513
		Mean(β_2)	0.000	0.020	1.729	1.740	1.030	1.048
$\rho_{x_1x_2}=0.5$	$\rho_{x_1y}=0.8$	Var(β_2)	0.240	0.244	0.330	0.362	0.291	0.296
		Mean(β_2)	0.000	0.022	1.472	1.492	1.007	1.036
$\rho_{x_1x_2}=0.8$	$\rho_{x_1y}=0.3$	Var(β_2)	1.264	1.295	0.784	0.742	0.934	0.931
		Mean(β_2)	0.000	0.029	2.453	2.476	1.678	1.692
$\rho_{x_1x_2}=0.8$	$\rho_{x_1y}=0.5$	Var(β_2)	1.042	1.055	0.678	0.679	0.788	0.824
		Mean(β_2)	0.000	0.013	3.925	3.908	2.684	2.678
$\rho_{x_1x_2}=0.8$	$\rho_{x_1y}=0.8$	Var(β_2)	0.500	0.460	0.420	0.426	0.432	0.495

5.2. *Efficiency*

Figure 2 shows the relative efficiency as measured by the residual sum of squares for the efficiency of the models, calculated from Eqs. (4) and (5). We can see that the categorized model always has a smaller residual sum of squares compared with the dichotomized model. But the efficiency of the model depends on the correlation between X_1 and Y. If the correlation between X_1 and Y is small to moderate (for example, < 0.5), there is not much loss in efficiency by discretizing X_1. Besides, there is not much difference between the dichotomized and categorical models. The difference between the two models increases when the correlation between X_1 and Y is high. In the extreme case when $\rho_{x_1y} = 0.9$, the relative of efficiency for the categorized model is almost twice that of the dichotomized model. The relative efficiency can be as low as around 0.4 for the dichotomized model in this case. The relative efficiency remains lower than 0.5 even though the correlation between X_1 and X_2 is as high as 0.7. In other words, adding X_2 does not rectify the problem. The relative efficiency for the categorized model is above 0.8 for all the cases in Fig. 2.

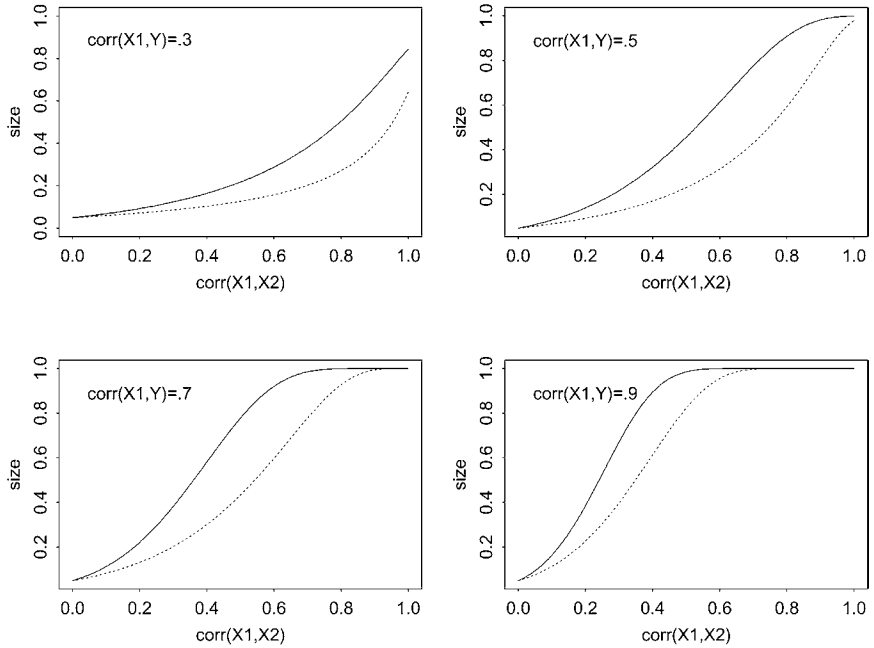

FIG. 1. Size of test of $\beta_2 = 0$ based on different models($n = 200$). (Solid line indicates the dichotomized model and dashed line the categorized model.)

Figure 3 illustrates the relative efficiency (6) of the general dichotomized model as a function of $p = P(X_1 < cutpoint)$ and ρ_{x_1y}. Here we fix $\rho_{x_1x_2} = 0.5$. From (6), we can see that the relative efficiency is a function of $1/\rho^2_{x_1x_2}$. So, using different $\rho_{x_1x_2}$ will only shift the curves up or down accordingly without changing the shape of the curves. Four different levels of ρ_{x_1y} are used in the graphs: 0.3, 0.5, 0.7, and 0.9. The relative efficiency is maximized when $p = 0.5$. As p deviates from 0.5, relative efficiency decreases. The curves are symmetric about $p = 0.5$. If p is near 0.5 the relative efficiency does not change much, particularly if the correlation between Y and X_1 is small, which means as long as we dichotomize the covariate near the median, we can still achieve similar relative efficiency.

6. CONCLUSIONS

In this paper, we demonstrate how categorizing a continuous variable can mislead one into concluding that a second unimportant variable is

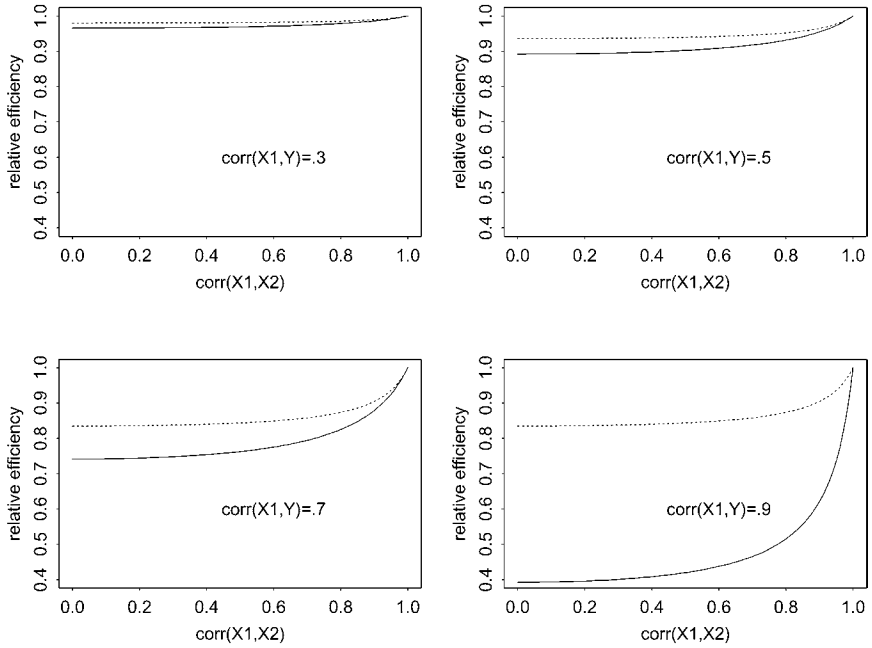

FIG. 2. Relative efficiency for prediction of dichotomized and categorized model compared to the continuous model. (Solid line indicates the dichotomized model and dashed line the categorized model.)

important and can lead to considerable loss of efficiency of prediction. The results are obtained assuming multivariate normality. The quantitative results would obviously differ for other distributional assumptions and other types of response variables. The algebraic expressions would also be quite different. However, we believe that qualitatively the results for other models will be strongly influenced by the correlation between the two explanatory variables and between the response and the discretized variable.

The unusual argument in favor of categorizing a continuous variable is because the functional form is not known and categorizing has some inherent robustness properties. In many applications, it is reasonable to think that the functional form of a covariate is continuous or even smooth and monotonic. Categorizing clearly does not lead to continuous or smooth functional forms. An alternative approach to categorizing, which acknowledges uncertainty in the functional form, is to estimate the functional form, along with the other parameters of interest, for example, by using a power transformation or nonparametric regression techniques.

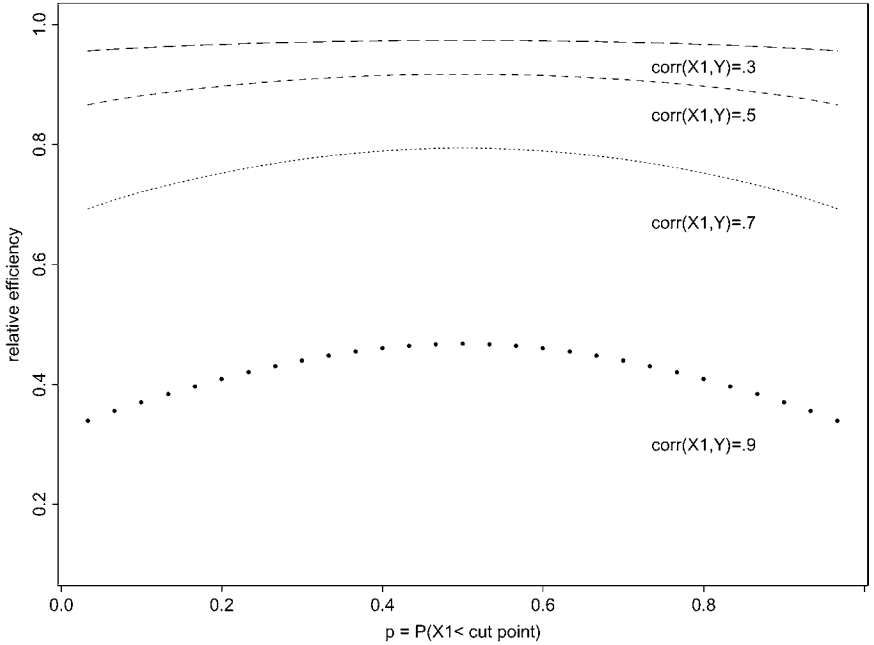

FIG. 3. Relative efficiency for prediction of dichotomizing a continuous covariate as a function of the cutpoint and correlation between Y and X_1, ($\rho_{x_1x_2} = 0.5$).

7. APPENDIX: DERIVATION OF BIAS AND VARIANCE.

We can express the denominator of $U(\mathbf{x}_1, \mathbf{x}_2)$ in Eq. (3) as

$$\frac{1}{n} s^2_{\mathbf{x}_1^*} \mathbf{X}_2' \left(\mathbf{I}_n - \frac{1}{n} \mathbf{J}_n - \frac{\widetilde{\mathbf{X}_1^*}' \widetilde{\mathbf{X}_1^*}}{n s^2_{\mathbf{X}_1^*}} \right) \mathbf{X}_2$$

where $\widetilde{\mathbf{X}_1^*}$ refers to the vector

$$\begin{pmatrix} X^*_{11} - \overline{X^*_1} \\ X^*_{12} - \overline{X^*_1} \\ \cdots \\ X^*_{1n} - \overline{X^*_1} \end{pmatrix}.$$

The matrix

$$\mathbf{I}_n - \frac{1}{n} \mathbf{J}_n - \frac{\widetilde{\mathbf{X}_1^*} \widetilde{\mathbf{X}_1^*}'}{n s^2_{\mathbf{X}_1^*}}$$

is idempotent, and its rank equals its trace:

$$n-1-tr\left(\frac{\widetilde{\mathbf{X}}_1^*\widetilde{\mathbf{X}}_1^{*\prime}}{ns_{X_1^*}^2}\right)=n-2.$$

The numerator of $E(\widehat{\beta_2^D} \mid X_1, X_2)$ is

$$\left[\frac{1}{n}s_{X_1^*}^2\widetilde{\mathbf{X}}_1' - \frac{1}{n}\,\widehat{\mathrm{cov}}(X_1^*, X_1)\,\widetilde{\mathbf{X}}_1^{*\prime}\right]\mathbf{X}_2\,\frac{\sigma_{\mathbf{x}_1 y}}{\sigma_{\mathbf{x}_1}^2}.$$

This has a normal distribution with mean 0 and variance

$$\begin{aligned}
&\frac{\sigma_{x_1 y}^2\sigma_{x_2}^2}{\sigma_{x_1}^4}\left[\frac{1}{n}s_{X_1^*}^2\widetilde{\mathbf{X}}_1' - \frac{1}{n}\,\widehat{\mathrm{cov}}(X_1^*, X_1)\,\widetilde{\mathbf{X}}_1^{*\prime}\right]\left[\frac{1}{n}s_{X_1^*}^2\tilde{\mathbf{X}}_1 - \frac{1}{n}\,\widehat{\mathrm{cov}}(X_1^*, X_1)\,\widetilde{\mathbf{X}}_1^{*}\right]\\
&\quad=\frac{\sigma_{\mathbf{x}_1 y}^2\sigma_{\mathbf{x}_2}^2}{\sigma_{\mathbf{x}_1}^4 n^2}\left[ns_{X_1^*}^4 s_{X_1}^2 - 2n\,\widehat{\mathrm{cov}}(X_1^*, X_1)^2 s_{X_1^*}^2 + ns_{X_1^*}^2\,\widehat{\mathrm{cov}}(X_1^*, X_1)^2\right]\\
&\quad=\frac{\sigma_{x_1 y}^2\sigma_{x_2}^2 s_{X_1^*}^2}{n\sigma_{x_1}^4}\left[s_{X_1^*}^2 s_{X_1}^2 - \widehat{\mathrm{cov}}(X_1^*, X_1)^2\right].
\end{aligned}$$

Furthermore the numerator is independent of the denominator given X_1 when n is large, since

$$\begin{aligned}
&\left[\frac{1}{n}s_{X_1^*}^2\widetilde{\mathbf{X}}_1' - \frac{1}{n}\,\widehat{\mathrm{cov}}(X_1^*, X_1)\,\widetilde{\mathbf{X}}_1^{*\prime}\right]\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n - \frac{\widetilde{\mathbf{X}}_1^*\widetilde{\mathbf{X}}_1^{*\prime}}{ns_{X_1^*}^2}\right)\\
&\quad=\frac{1}{n}s_{X_1^*}^2\widetilde{\mathbf{X}}_1'\left(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_n - \frac{\widetilde{\mathbf{X}}_1^*\widetilde{\mathbf{X}}_1^{*\prime}}{ns_{X_1^*}^2}\right)\\
&\quad=\frac{1}{n^2}\widetilde{\mathbf{X}}_1'\widetilde{\mathbf{X}}_1^*\widetilde{\mathbf{X}}_1^{*\prime}\\
&\quad=\frac{1}{n}\,\widehat{\mathrm{cov}}(X_1^*, X_1)\,\widetilde{\mathbf{X}}_1^{*\prime}\\
&\quad=O\left(\frac{1}{n}\right).
\end{aligned}$$

Thus for large n

$$\begin{aligned}
U(\mathbf{X}_1, \mathbf{X}_2) \mid \mathbf{X}_1 = \mathbf{x}_1 &\sim \frac{\frac{\sigma_{x_1 y}\sigma_{x_2}s_{X_1^*}}{\sqrt{n}\,\sigma_{x_1}^2}\sqrt{\left[s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_1}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, x_1)^2\right]}\,\mathbf{N}(0,1)}{\frac{s_{\mathbf{x}_1^*}^2\cdot\sigma_{x_2}^2}{n}\cdot\chi_{n-2}^2}\\
&\equiv\frac{\sqrt{n}\,\sigma_{x_1 y}\sqrt{\left[s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_1}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_1)^2\right]}}{s_{\mathbf{x}_1^*}\sigma_{x_1}^2\sigma_{x_2}^2}\frac{\mathbf{N}(0,1)}{\chi_{n-2}^2}.
\end{aligned}$$

The unconditional variance of $U(\mathbf{X}_1, \mathbf{X}_2)$ is given by

$$\begin{aligned}
\mathrm{Var}[U(\mathbf{X}_1, \mathbf{X}_2)] &= E_{\mathbf{X}_1}\{\mathrm{Var}_{\mathbf{X}_2}[U(\mathbf{X}_1, \mathbf{X}_2)]\} + \mathrm{Var}_{\mathbf{X}_1}\{E_{X_2}[U(\mathbf{X}_1, \mathbf{X}_2)]\} \\
&= E\left[\frac{n\sigma_{x_1y}^2 \cdot [s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_1}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_1)^2]}{s_{\mathbf{x}_1^*}{}^2 \sigma_{x_1}{}^4 \sigma_{x_2}{}^2}\right] \cdot \mathrm{Var}\left(\frac{\mathbf{N}(0,1)}{\chi_{n-2}^2}\right) + 0 \\
&= \frac{n\sigma_{x_1y}^2}{\sigma_{x_1}^4 \sigma_{x_2}^2} \cdot E\left[\frac{s_{\mathbf{x}_1^*}^2 s_{\mathbf{x}_1}^2 - \widehat{\mathrm{cov}}(\mathbf{x}_1^*, \mathbf{x}_1)^2}{s_{\mathbf{x}_1^*}^2}\right] \cdot E\left[\left(\frac{\mathbf{N}(0,1)}{\chi_{n-2}^2}\right)^2\right] \\
&= \frac{n\sigma_{x_1y}^2}{\sigma_{x_1}^4 \sigma_{x_2}^2}\left(\sigma_{x_1}^2 - \frac{2}{\pi}\sigma_{x_1}^2\right) E\left[\left(\frac{\mathbf{N}(0,1)}{\chi_{n-2}^2}\right)^2\right] \\
&= \frac{0.3634\rho_{x_1y}^2 \sigma_y^2}{n\sigma_{x_2}^2}.
\end{aligned}$$

Similarly we can show that

$$E[V(\mathbf{X}_1, \mathbf{X}_2)] = \frac{1-\rho_{x_1y}^2}{1-0.6366\rho_{x_1x_2}^2} \cdot \frac{\sigma_y^2}{n\sigma_{x_2}^2}.$$

REFERENCES

1. D. G. Altman, Categorising continuous variables, *British J. Cancer* **64** (1991), 975.
2. D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher, Dangers of using optimal cutpoints in the evaluation of prognostic factors, *J. National Cancer Institute* **86** (1994), 829–835.
3. R. J. Connor, Grouping for testing trends in categorical data, *J. Amer. Statist. Assoc.* **67** (1972), 601–604.
4. D. R. Cox, Note on grouping, *J. Amer. Statist. Assoc.* **52** (1957), 543–547.
5. F. A. Graybill, "Theory and Application of the Linear Model," Duxbury Press, London, 1976.
6. F. E. Harrell, Jr., K. L. Lee, and D. B. Mark, Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statist. Medicine* **15** (1996), 367–387.
7. S. W. Lagakos, Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable, *Statist. Medicine* **7** (1988), 257–274.
8. B. Lausen and W. Schumacher, Evaluating the effect of optimized cutoff values in the assessment of prognostic factors, *Comp. Statist. Data Anal.* **21** (1996), 307–326.
9. T. M. Morgan and R. M. Elashoff, Effect of categorizing a continuous covariate on the comparison of survival time, *J. Amer. Statist. Assoc.* **81** (1986), 917–921.
10. L. P. Zhao and L. N. Kolonel, Efficiency loss from categorizing quantitative exposures into qualitative exposures in case-control studies, *Amer. J. Epidemiology* **136** (1992), 464–474.

Printed in The Netherlands