# Exploring the Issues of Truncation of Continuous Variables as Covariates in Linear Regression using OKLAHOMA BRFSS Data

Authors: Siew Ang[1], MA, Derek Pate[2], MPH, Jorge Mendoza[3], PhD
[1, 2] Oklahoma State Department of Health, [3] University of Oklahoma, Psychology Department

## Background

Covariates are commonly used when there are confounding factors in a study and these factors have not been incorporated into the design of the study. In most statistical textbooks, covariates are usually recommended as continuous variables. However, it is common in social and medical studies to truncate, dichotomize, or categorize a continuous variable despite warnings in the literature of problems with such practices. These practices may be problematic when categorized variables are treated as covariates, leading to potential bias in the regression estimates and faulty conclusions.

## Objectives

The objective of this study is to explore and compare the effects of truncated variables either as criterion variable or covariates in sets of simple linear regressions, (i.e. age and exercise on weight), and to examine the influence of measurement errors on these effects. This study is also an attempt to bridge the gap between the literature and practical field by demonstrating how under certain circumstances, a continuous covariate might yield a more precise result than a categorized continuous variable.

## Methods

We utilized four variables from the 2005 Oklahoma Behavioral Risk Factor Surveillance System (BRFSS) data to demonstrate our objectives:

- LTPA – whether respondents had participated in ANY Leisure time physical activity in the past month – consisted of dichotomous Yes / No responses.
- AGE – age of respondents (in years).
- BMI – body mass index was computed as weight in kilograms divided by height in meters squared, (weight/height**2).
- BMICAT – computed from BMI, consisted of three categories: normal, overweight and obese groups. Adults with BMI < 25.00 were classified as normal weight, 25.00 <= BMI < 30.00 as overweight, 30.00 <= BMI < 99.99 as obese.

AGE was chosen as the covariate in this study because: 1) the measurement error of AGE is generally small, and thus provided a necessary baseline for comparison after inducing measurement errors, and 2) AGE was commonly known to influence exercise levels. Continuous AGE was induced with random measurement errors (SD = 2, 4), named as AGE2SD and AGE4SD (reliability = 0.96, 0.57). Measurement errors SD = 0, 2 were defined as small measurement errors, while SD = 4 was defined as large measurement error or imprecise measurement. AGE, AGE2SD and AGE4SD were grouped in two ways: 1) dichotomizing age at various thresholds: at > 35, 40, 45, 50, and 55 years old; the first three thresholds were below the average age, while the last two thresholds were above the mean age; and 2) age regrouped into 2, 3, 7 and 15 groups. Two groups of AGE (2 GRPS) were split at threshold 45. Three groups (3 GRPS) were formed at $\leq$ 34, 35-64, 65+. Seven groups (7 GRPS) started at age $\leq$ 24, and then grouped every ten years thereafter. Fifteen groups (15 GRPS) were categorized at age $\leq$ 20 and every five years.

The relationships between BMI, BMICAT and LTPA with 1) dichotomized AGE and 2) categorized AGE, were first explored using SAS Proc Corr for Pearson correlation coefficients (Rhos). As BRFSS data require an adjusting/ weighting procedure to statistically reflect the parameters for the population, a "weight" subcommand was included in all the procedures reported here. Pearson correlation is a test statistic commonly used to test the relationship between two continuous variables. When one variable is dichotomized or categorized, Pearson correlation coefficient naturally becomes point-biserial correlation. A Rho of $\geq$ 0.3 between covariate-predictor indicates a moderate correlation. For comparison purposes, this study included the Rhos for the continuous AGE, AGE2SD and AGE4SD in each of the figures as a baseline (red line).

We performed a series of simple linear regressions using SAS Proc Surveyfreq with a "weight" subcommand to reflect the population values. The full regression model was:

$$Y = \beta_0 + \beta_1 * AGE + \beta_2 * LTPA + e$$

in which

Y is the criterion variable, BMI or BMICAT
$\beta_0$ (Beta Naught) is the regression coefficient for the intercept
$\beta_1$ (Beta One) is the regression coefficient or the slope for AGE
$\beta_2$ (Beta Two) is the regression coefficient or the slope for LTPA
E is the error term

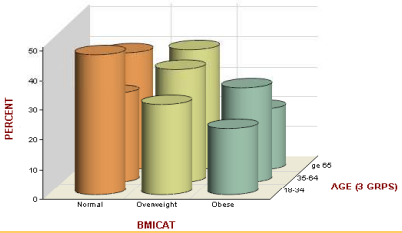
Table 1

Sample sizes (N), Averages (Mean), Standard Deviations (SD) and 95% Confidence Intervals (CI) for BMI, AGE, AGE2SD, AGE4SD, Oklahoma 2005

| Variable | Label | N | Mean | SD | 95% CI |
|---|---|---|---|---|---|
| BMI | BODY MASS INDEX | 13158 | 27.37 | 81.01 | 27.27 - 27.47 |
| AGE | REPORTED AGE (SD = 0) | 13673 | 45.93 | 253.00 | 45.63 - 46.24 |
| AGE2SD | INDUCED AGE (SD = 2) | 13673 | 45.97 | 258.72 | 45.65 - 46.28 |
| AGE4SD | INDUCED AGE (SD = 4) | 13673 | 46.08 | 335.79 | 45.67 - 46.48 |

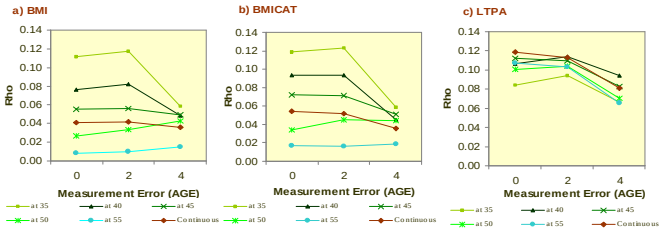
Figure 1

Proportions of Adults in Weight Categories by Age, Oklahoma 2005



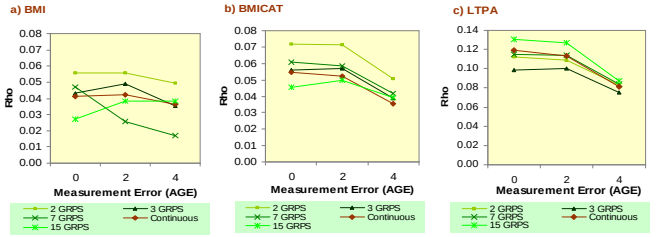

Figures 2a-c

Effects of Dichotomizing AGE and its Relationships with BMI, BMICAT and LTPA as Measurement Errors on AGE Increase



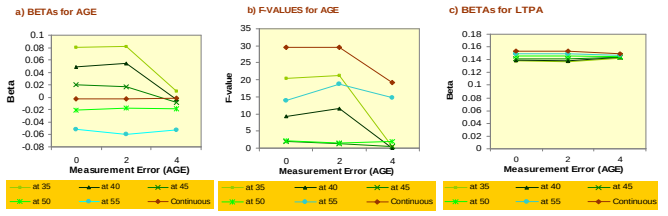

Figures 3a-c

Effects of Truncating AGE and Its Relationships with BMI, BMICAT and LTPA as Measurement Errors on AGE Increase



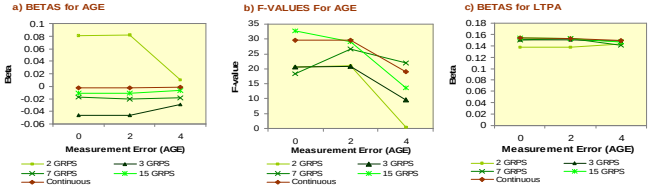

Figures 4a-c

Effects of Dichotomization on Linear Regression Results for AGE and LTPA on BMICAT as Measurement Errors on AGE increase




Figures 5a-c

Effects of Truncation on Linear Regression Results for AGE and LTPA on BMICAT as Measurement Errors on AGE Increase



Two models were performed: the restricted and the full models. In the restricted model, AGE was the only variable included. The slope of AGE was tested with a F-statistic. If the F-value was significant at $p < 0.05$, then the full model was performed. The Betas for LTPA and AGE, and the F-values for the Betas of AGE were plotted.

**Results**

• Table 1 shows the N sizes, means, standard deviations and 95% confidence intervals for BMI, AGE, AGE2SD and AGE4SD.

• Figure 1 illustrates the weighted prevalence of BMICAT for Oklahoma adults by Age 3 GRPS in 2005. There was a strong association between AGE and BMICAT ($\chi^2 = 139.69$, $p < 0.0001$). Being overweight and obese was more prevalent among adults aged 35 and above. Association was also found between AGE 3 GRPS and LTPA ($\chi^2 = 41.98$, $p < 0.0001$).

• Figures 2a-c compare the Rhos of AGE with BMI, BMICAT and LTPA. As the sample sizes were very large, all the Rhos were statistically significant and the p-values will not be reported here.
• In Figure 2a, when measurement error = 0, the Rhos between binary AGE and BMI changed according to the threshold in which AGE was dichotomized. When AGE was dichotomized at 35, 40 and 45 (below the mean AGE), the Rhos were inflated. However, when AGE was dichotomized at 50 and 55, the Rhos were underestimated. As imprecision increased to SD = 2, Rhos were inflated at all thresholds. As imprecision increased to SD = 4, Rhos (for thresholds below the mean AGE) dropped, while Rhos (for thresholds above the mean AGE) continued to increase.
•Figure 2b shows the same pattern as Figure 2a when both variables were truncated (BMICAT and binary AGE).
• Figure 2c illustrates that the fluctuation of correlations between a dichotomized variable and a naturally occurring binary variable were relatively small. Under measurement error = 0, the Rhos ranged only from 0.08 to 0.12. Dichotomizing AGE at threshold 45 had a Rho closest to the baseline. When imprecision for AGE increased to SD = 2, the Rhos were slightly elevated for thresholds 35 and 40, while the other Rhos decreased. However, as imprecision further increased to SD = 4, all the Rhos decreased. Threshold at 45 again had the closest Rho compared to the baseline.

• Figures 3a-c compare the effects of multiple truncations on correlations of AGE with BMI, BMICAT and LTPA.
• Figure 3a demonstrates that truncation of continuous AGE variable into 2 or more groups affected Rhos. At measurement error = 0, 2 GRPS had the highest Rho inflation. The Rho for 3 GRPS was closest to the Rho for the baseline. The Rho for 15 GRPS was underestimated compared to the baseline. As imprecision increased to SD = 2, the Rhos for 3 GRPS and 15 GRPS were inflated, whereas the Rho for 7 GRPS was underestimated. As measurement error increased to SD = 4, the Rhos for 3 GRPS and 15 GRPS were closest to the baseline, while 7 GRPS continued to drop. Of all the truncation methods, only the Rho for 15 GRPS increased as imprecision increased. The reason for the increase was unknown, but was speculated to be associated with truncation.
• When another truncated variable (BMICAT) was correlated with these truncated AGE groups (in Figure 3b), the patterns became more predictable. As imprecision increased to SD= 4, all the Rho values decreased.
• Trends similar to Figure 3b were exhibited in Figure 3c as measurement errors increased.

• Figures 4a-c demonstrate the regression results for dichotomized AGE and LTPA on the criterion variable BMICAT. As all the F-values for LTPA were highly significant under all scenarios ($p < 0.001$), they were not included in these figures. • Figure 4a shows the Betas for dichotomized AGE as covariates after regressing on BMICAT as compared to the Beta for the baseline. The Betas, under measurement error = 0, were overestimated for AGE dichotomized at thresholds 35, 40 and 45, while the Betas for thresholds 50 and 55 were underestimated. As imprecision increased to SD = 2, the Betas for binary age at 35 and 40 were inflated, and Betas for threshold at 55 was deflated. However, as imprecision continued to increase to SD = 4, the Beta values for dichotomized AGE became closer to the baseline (became closer to zero).

• Figure 4b illustrates that F-values were mostly compromised when AGE was dichotomized as compared to the baseline under no measurement error. However, as imprecision increased to SD = 4, all the F-values decreased including the baseline.
• Figure 4c indicates that the Betas for LTPA remained relatively close to each other despite of the various thresholds for AGE covariates when measurement error = 0. As imprecision increased to SD = 4, the Betas for LTPA converge to the baseline.

• Figures 5a-c demonstrate the regression results for multiple truncated AGE and LTPA on criterion variable BMICAT. F-values for LTPA were highly significant under all scenarios ($p < 0.001$), therefore, they were not included in these figures.
• In Figure 5a, the Beta for 2 GRPS was overestimated, while 3 GRPS, 7 GRPS and 15 GRPS were underestimated under no measurement error scenario. The Beta for 15 GRPS was closest to the baseline. As imprecision increased to SD = 4, all the Betas became closer to the baseline.
• Figure 5b indicates that except for 15 GRPS, all the F-values for truncated AGE were underestimated at measurement error = 0. As imprecision increased to SD= 4, all the F-values decreased.
Figure 5c illustrates that Betas for LTPA were relatively unaffected in the presence of truncated AGE groups, even as measurement errors increased.

**Conclusions and Discussion**

• This study demonstrates the possible pitfalls of simply dichotomizing continuous variables when collected as continuous. The thresholds for dichotomization and the number of truncation affected the relationships between variables in a specific manner.
• Under no measurement errors, dichotomizing AGE at thresholds below the mean AGE tended to inflate the correlations with a continuous BMI or a truncated variable such as BMICAT, while choosing the thresholds above the mean were more likely to decrease these correlations. However, dichotomizing AGE was more likely to underestimate correlations with a naturally occurring binary variable such as LTPA.
• The higher the number of truncations for continuous AGE, the more likely it was for the results to show similar patterns of the baseline.
• Truncating or dichotomizing continuous criterion variable into BMICAT, although exhibiting similar patterns for the Betas and F-values as for BMI, changed the scales of coefficients, making the scale smaller and more difficult to interpret.
• When AGE was imprecise (large measurement error), dichotomizing or truncating AGE tended to converge the correlations and Betas towards the baseline values.
• When the correlations for AGE-LTPA were small (the highest value was 0.12), it was more difficult to demonstrate a large effect of dichotomizing or truncating AGE on the LTPA. Had the correlations for covariate-predictor been larger, the Betas and F-values for the predictor would be greatly affected.
• If the correlations between variables were larger (greater than .3), we would be able to increase the measurement errors until the correlations tapered down to zero.

**Recommendations for Future Studies**

•Future study should explore the effects of truncation when covariate-predictor relationship is very strong, say, $r \geq 0.6$ by systematically examine the criterion-covariate-predictor relationships in simulation studies.
• Future study should also examine the effects of systematic measurement error and these relationships in logistic regression settings.

**For Further Question or Comment, Please Contact:**

Siew C. Ang
Oklahoma State Department of Health,
Health Care Information Division
1000 NE 10th Street Rm 807
OKC, OK 73117
Tel: (405) 271-9444 x 56473
Email: siewa@health.ok.gov