

1 Brief methods note: Investigators must exercise caution prior  
2 to dichotomizing continuous variables for use as binary  
3 outcomes in logistic regression

4 James Uanhoro<sup>1</sup>

5 <sup>1</sup> The Ohio State University

6 Abstract

7 A common practice in empirical data analysis is to dichotomize continuous outcomes for substantive or interpretational purposes. A particular cut-point on the outcome variable may be practically relevant, such that researchers dichotomize the continuous outcome at the cut-point to create a binary outcome, then proceed to model this binary outcome using logistic regression. However, homoskedasticity of the error term in the linear regression model for the continuous outcome is an often overlooked assumption for this application of logistic regression to be valid. If this condition is not met, the logistic regression model for the binary outcome will be misspecified, and the coefficients and predicted probabilities will be incorrect. Correctly estimating the relationship can be difficult computationally. Hence, I recommend that researchers directly model the continuous outcome even when there are substantive justifications for modeling the dichotomized outcome.

*Keywords:* logistic regression, dichotomization, heteroskedasticity

8 **Introduction**

9 It is a relatively common practice for education researchers to dichotomize continuous  
10 variables into binary variables, then analyze the binary variables using logistic regression.  
11 The most common reason for doing this is substantive. For example, the original response  
12 variable may be student scores on a test with range of 0 to 100. On this test, students with  
13 scores equal to or above 70 pass, while others fail. In such situations, researchers interested  
14 in understanding the dynamics of student success on the test may dichotomize the original

---

James Uanhoro is a PhD student in the Quantitative Research, Evaluation and Measurement program within the Department of Educational Studies

Correspondence concerning this article should be addressed to James Uanhoro, 210 Ramseyer Hall, 29 W Woodruff Ave, Columbus, OH 43210. E-mail: uanhoro.1@osu.edu

15 test scores into 1 (pass) or 0 (fail) and model the probability that a student passed (or failed)  
 16 using logistic regression.

17 In this paper, I argue that there are reasons to distrust the findings from such a  
 18 procedure. In the next section, I review two formulations for the logistic regression model.  
 19 By doing this, it becomes evident why this procedure can be problematic. The source of  
 20 this problem is heteroskedasticity in the linear regression model for the original continuous  
 21 variable. I conclude with a simple demonstraton using simulated data.

## 22 Derivation of the logistic regression model

23 The standard logistic regression model is a generalized linear model (Fox, 2015;  
 24 McCullagh & Nelder, 1989) for the probabilities responsible for an observed Bernoulli  
 25 response variable:

$$\ln \left[ \frac{p}{1-p} \right] = \alpha + X\beta \quad (1)$$

26 where  $p$  are the probabilities of success underlying the observed Bernoulli variable for  
 27 each case,  $\alpha$  is the intercept,  $X$  is an  $n$  by  $k$  matrix for  $n$  cases and  $k$  predictors (excluding  
 28 the intercept), and  $\beta$  are  $k$  regression weights for each predictor. The logit transformation  
 29 applied to the probabilities (left hand side of equation (1)) equates the probabilities to the  
 30 predictors,  $X$ , multiplied by their weights,  $\beta$ , plus the intercept,  $\alpha$ .

31 We can rewrite equation (1) by solving for  $p$ :

$$p = \frac{e^{(\alpha+X\beta)}}{1 + e^{(\alpha+X\beta)}} \quad (2)$$

32 Equation (2) is also known as the inverse logit transformation applied to  $\alpha + X\beta$ ,  
 33 transforming it from a value that has a possible range of  $-\infty$  to  $\infty$  to a probability guaranteed  
 34 to lie between 0 and 1. At this point, I motivate the logistic regression model using a latent  
 35 variable formulation (Amemiya, 1981):

$$p = P(\alpha + X\beta + \epsilon > t) \quad (3)$$

36 This formulation of the model is the one we rely on when we dichotomize a continuous  
 37 variable for use as the outcome in a logistic regression model. The probabilities underlying  
 38 the new binary outcome are the probabilities that a continuous variable with a systematic  
 39 component,  $\alpha + X\beta$ , and random error,  $\epsilon$ , exceeds a threshold,  $t$ . For logistic regression,  
 40 we make the additional assumption that  $\epsilon$  is a standard logistic variable ( $\epsilon \sim \mathcal{L}(0, 1)$ ); this  
 41 means that  $\epsilon$  has mean 0 and variance of  $\pi^2/3$ .

42 As one will observe from equation (3), changing the value of  $t$  simply changes the  
 43 value of  $\alpha$ . If the threshold increases by 3, then the intercept increases by 3. So for the

44 model to be identified, we will assume the threshold is 0,  $t = 0$ . Given this information, we  
 45 continue to solve for  $p$  in equation (3):

$$\begin{aligned} p &= P(\alpha + X\beta + \epsilon > 0) = P(\epsilon > -\alpha - X\beta) \\ &= P(\epsilon < \alpha + X\beta) \quad \text{since } \mathcal{L}(0, 1) \text{ is symmetric about } 0 \end{aligned} \quad (4)$$

46 The final line of equation (4) is simply the cumulative distribution function of  $\mathcal{L}(0, 1)$   
 47 evaluated at  $\alpha + X\beta$ , hence:

$$p = \frac{e^{(\alpha + X\beta)}}{1 + e^{(\alpha + X\beta)}} \quad (5)$$

48 This result in equation (5) shows that the latent variable formulation for the logistic  
 49 regression model is equivalent to the generalized linear model formulation for logistic  
 50 regression in equation (2). It also reveals one major assumption for the standard logistic  
 51 regression model to be valid: the random error must be homoskedastic i.e.  $\epsilon \sim \mathcal{L}(0, 1)$ . If  
 52 this assumption is violated, then equation (5) is wrong. Assuming  $\epsilon \sim \mathcal{L}(0, \sigma)$  instead where  
 53  $\sigma$  has a different value for each case, the correct equation is:

$$p = \frac{e^{\left(\frac{\alpha + X\beta}{\sigma}\right)}}{1 + e^{\left(\frac{\alpha + X\beta}{\sigma}\right)}} \quad (6)$$

54 Hence if the random error is heteroskedastic, the standard logistic regression model as  
 55 implemented in statistical software packages will be inadequate if the model is applied to  
 56 the dichotomized outcome.

### 57 **How problematic can this form of heteroskedasticity be?**

58 To illustrate the problem, I present a simple example. Assume the following regression  
 59 equation for a continuous variable,  $z_i$ :  $z_i = x_i + \epsilon_i$ , where  $i = 1, 2, \dots, 5000$ ,  $x_i \sim \mathcal{N}(0, 1)$   
 60 and  $\epsilon_i \sim \mathcal{L}(0, 0.5 + \frac{1}{4}\sqrt{e^{x_i}})$ , so the error variance is non-constant and depends on  $x_i$ . I  
 61 dichotomize  $z_i$  at 0 to create a new binary response,  $y_i$ , such that  $y_i = 1$  where  $z_i > 0$   
 62 and  $y_i = 0$  when  $z_i \leq 0$ . So  $z_i$  is the continuous variable underlying the binary outcome,  
 63  $y_i$ . For comparison, I also generated a homoskedastic  $z_i$  and  $y_i$  under the condition that  
 64  $\epsilon_i \sim \mathcal{L}(0, 1)$ . As is visible from Figure 1, the average relationship between  $x_i$  and each  $z_i$  is  
 65 not that different under homoskedasticity and heteroskedasticity, but the heteroskedastic  $z_i$   
 66 visibly displays non-constant error variance.

67 Next, I regressed the continuous variable,  $z_i$ , on  $x_i$  using linear regression. Regardless  
 68 of the error variance structure, the linear regression model had a decent recovery of the  
 69 coefficient for  $x_i$ ; both coefficients were quite close to 1 (see first two columns of Table 1).  
 70 This is consistent with the theory on linear regression models: unbiased coefficient estimation

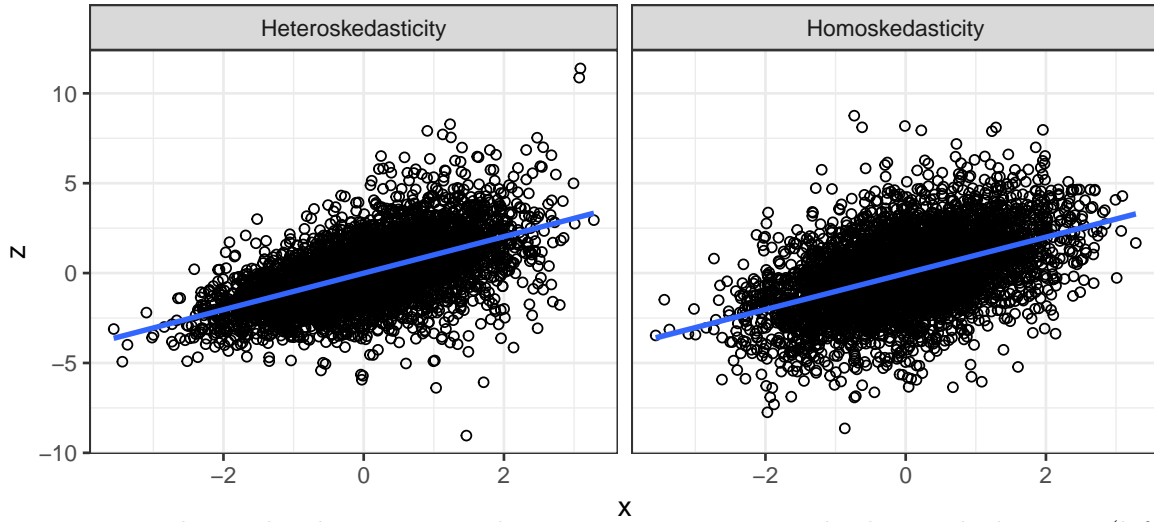


Figure 1. Relationship between  $x$  and continuous outcome under heteroskedasticity (left panel) and homoskedasticity (right panel). Under heteroskedasticity, the variance of  $z$  is larger at higher values of  $x$ .

Table 1  
Regression of outcome variables on  $x$

	Dependent variable:			
	Homoskedastic $z$	Heteroskedastic $z$	Homoskedastic $y$	Heteroskedastic $y$
$x$	1.007*** (0.026)	1.016*** (0.021)	1.027*** (0.038)	1.316*** (0.043)
Constant	-0.014 (0.025)	-0.004 (0.021)	-0.031 (0.031)	-0.175*** (0.033)
Observations	5,000	5,000	5,000	5,000

Note: \* $p < 0.05$ ; \*\* $p < 0.01$ ; \*\*\* $p < 0.001$   
The models for  $z$  are linear regression models. The models for  $y$  are logistic regression models.

71 does not dependent on assumptions like homoskedasticity or normality of errors (Gelman &  
72 Hill, 2007, p. 46).

73 I next regressed the binary variable,  $y_i$ , on  $x_i$  using logistic regression. Under ho-  
74 moskedastic error variance, the coefficient of  $x_i$  was 1.027, close to 1. However, under  
75 heteroskedastic error variance, the coefficient of  $x_i$  was about 30% higher  $((1.316 - 1) \times 100\%)$   
76 than what I specified in the data generation process. In this situation, we have inflation of  
77 the coefficient. Depending on the form of heteroskedasticity, the result might be coefficient  
78 deflation.

79 This happens because the standard logistic regression model is misspecified under  
80 heteroskedasticity. And one cannot recover the true logistic regression coefficients using  
81 standard maximum likelihood estimation of the logistic regression model.

## 82 Discussion

83 If the functional form of heteroskedasticity is known, then it is possible to modify  
84 the likelihood function used in maximum likelihood estimation of the logistic regression  
85 model. The literature is more developed in the case of probit models, where there exists a  
86 class of models known as *heteroskedastic probit models* (Alvarez & Brehm, 1995). However,  
87 even if the researcher can identify the functional form for heteroskedasticity, there is no  
88 guarantee that estimating the model will result in a correct solution. Potential problems  
89 include multiple solutions with near equivalent fit to the data, near singular Hessian matrices,  
90 large standard errors and convergence failures (Keele & Park, 2006).

91 Hence, I recommend that when researchers have access to the original continuous  
92 variable, they should model this variable regardless of questions of substantive interest. If  
93 education researchers are interested in studying relationships at thresholds that are very  
94 different from the mean of the outcome, quantile regression (Koenker & Hallock, 2001) is  
95 one approach for exploring the relationship between the predictors and continuous outcome  
96 at different quantiles of the outcome. There are methods for converting linear regression  
97 coefficients to logits and odds ratios (Moser & Coombs, 2004) but they also rely on the  
98 aforementioned homoskedasticity assumption.

99 Finally, the problem of heteroscedasticity described above can exist even when the  
100 outcome variable is truly binary, or the binary outcome is difficult to rationalize as the  
101 manifestation of a dichotomized continuous variable. I focus on the case where the investigator  
102 has access to the continuous variable here because the situation is readily salvageable: analyze  
103 the continuous outcome directly. In situations where the investigator does not have access  
104 to the underlying continuous variable, but heteroskedasticity may be a concern, more  
105 flexible regression approaches such as generalized additive models (Hastie, 2017) and kernel  
106 regularized least squares (with logistic loss, Hainmueller & Hazlett, 2014) may yield results  
107 that are more likely to reflect the true relations in the data.

## References

108

- 109 Alvarez, R. M., & Brehm, J. (1995). American ambivalence towards abortion policy:  
110 Development of a heteroskedastic probit model of competing values. *American*  
111 *Journal of Political Science*, 1055–1082.
- 112 Amemiya, T. (1981). Qualitative Response Models: A Survey. *Journal of Economic*  
113 *Literature*, 19(4), 1483–1536. doi:10.2307/2724565
- 114 Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.
- 115 Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical*  
116 *models* (p. 625). Cambridge University Press.
- 117 Hainmueller, J., & Hazlett, C. (2014). Kernel regularized least squares: Reducing misspecifi-  
118 cation bias with a flexible and interpretable machine learning approach. *Political*  
119 *Analysis*, 22(2), 143–168. doi:10.1093/pan/mpt019
- 120 Hastie, T. J. (2017). Generalized additive models. In *Statistical models in s* (pp. 249–307).  
121 Routledge.
- 122 Keele, L., & Park, D. K. (2006). *Ambivalent about ambivalence: A re-examination of*  
123 *heteroskedastic probit models*. Unpublished manuscript.
- 124 Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*,  
125 15(4), 143–156.
- 126 McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models*. Chapman & Hall/CRC.
- 127 Moser, B. K., & Coombs, L. P. (2004). Odds ratios for a continuous outcome variable without  
128 dichotomizing. *Statistics in Medicine*, 23(12), 1843–1860. doi:10.1002/sim.1776