IPS / Translational Bioinformatics

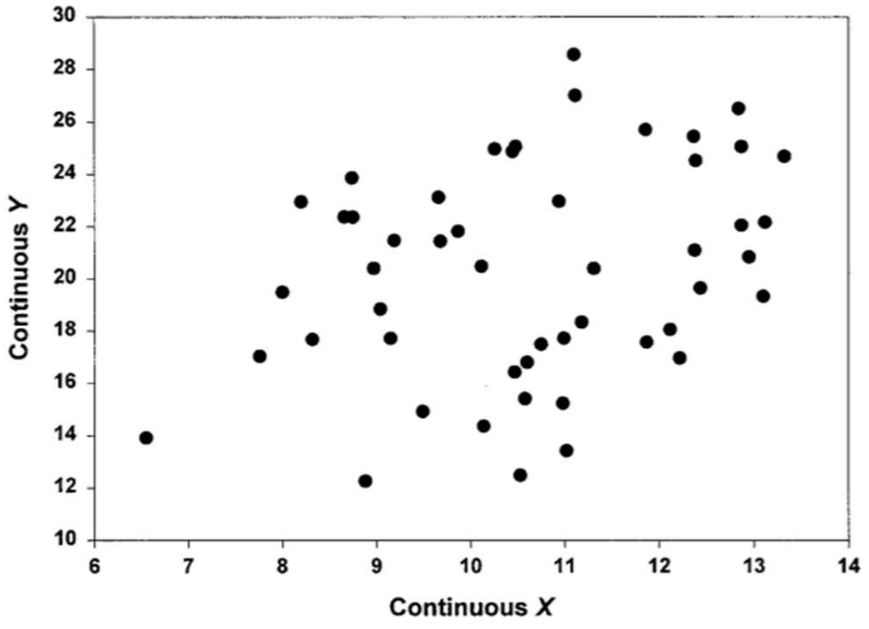
Ending Dichotomania: Stop Throwing Away Information

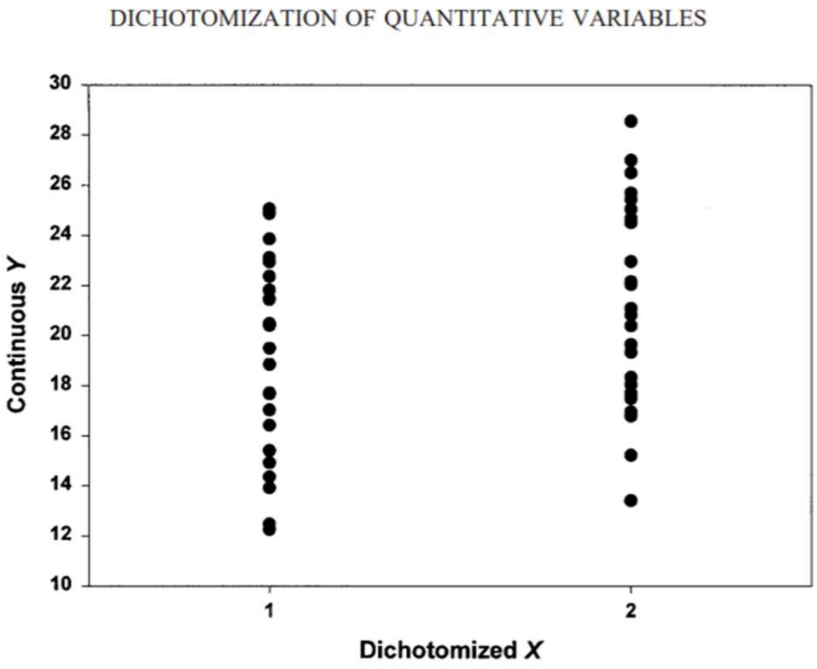
September 23, 2021

Abraham Apfel and Scott Chasalow
IPS/BIOM

Bristol Myers Squibb™

What Does it Mean to Dichotomize?





Figures copied from:

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Different Relationships on Continuous Scale Can Yield Identical Relationship After Dichotomization

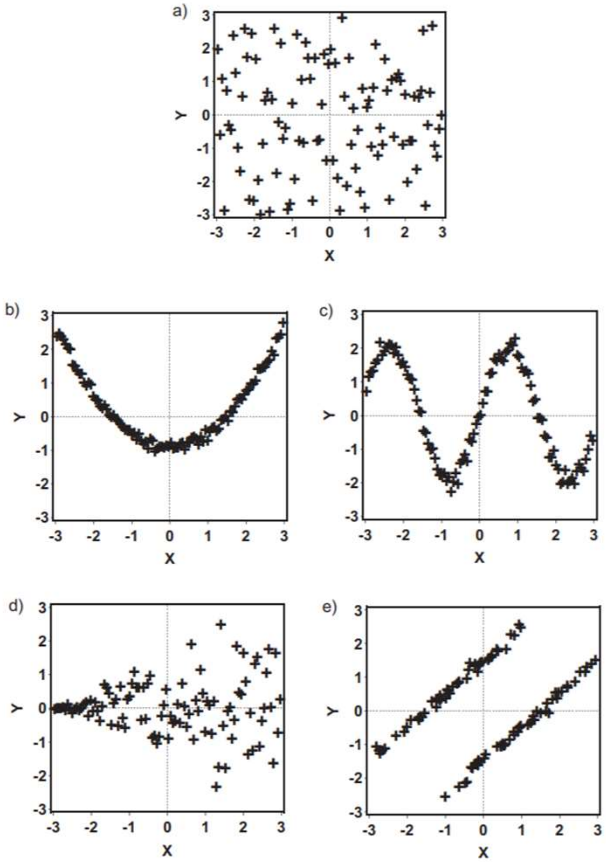


Fig. 1. Five scatterplots that yield an identical fourfold table (see table 1) when X and Y are both dichotomized at the value 0

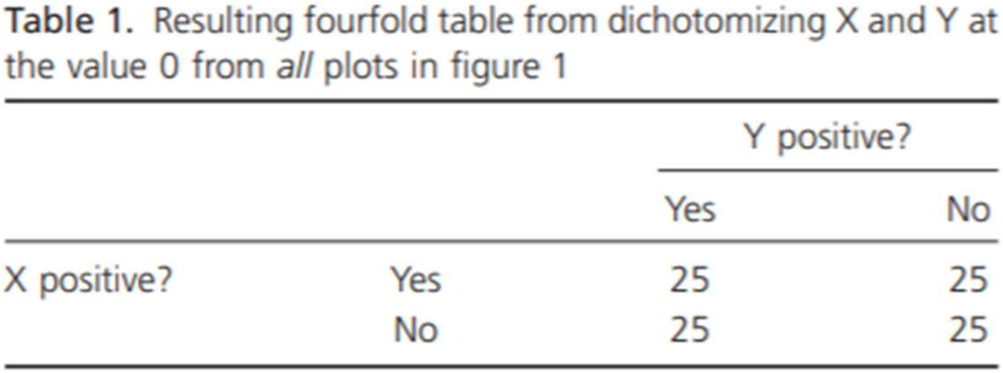


Table 1. Resulting fourfold table from dichotomizing X and Y at the value 0 from *all* plots in figure 1

		Y positive?	
		Yes	No
X positive?	Yes	25	25
	No	25	25

Table and Figure copied from:

Kuss, Oliver. "The danger of dichotomizing continuous variables: A visualization." *Teaching Statistics* 35.2 (2013): 78-79.

Ground-breaking? Hardly!

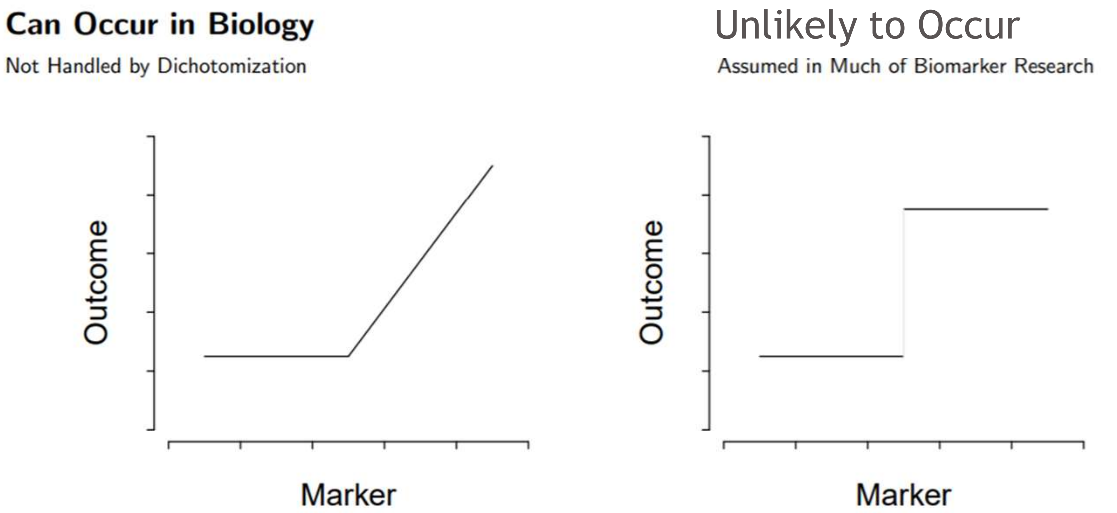
1. Cox, David R. "Note on grouping." *Journal of the American Statistical Association* 52.280 (**1957**): 543-547.
2. Senn, Stephen. "Dichotomania: an obsessive compulsive disorder that is badly affecting the quality of analysis of pharmaceutical trials." *Proceedings of the International Statistical Institute, 55th Session, Sydney* (2005).
3. Altman, Douglas G., and Patrick Royston. "The cost of dichotomising continuous variables." *Bmj* 332.7549 (2006): 1080.
4. Altman, Douglas G. "Suboptimal analysis using 'optimal' cutpoints." *British journal of cancer* 78.4 (1998): 556.
5. Royston, Patrick, Douglas G. Altman, and Willi Sauerbrei. "Dichotomizing continuous predictors in multiple regression: a bad idea." *Statistics in medicine* 25.1 (2006): 127-141.
6. Buettner, Petra, Claus Garbe, and Irene Guggenmoos-Holzmann. "Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma." *Journal of clinical epidemiology* 50.11 (1997): 1201-1210.
7. Fedorov, Valerii, Frank Mannino, and Rongmei Zhang. "Consequences of dichotomization." *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 8.1 (2009): 50-61.
8. Giannoni, Alberto, et al. "Do optimal prognostic thresholds in continuous physiological variables really exist? Analysis of origin of apparent thresholds, with systematic review for peak oxygen consumption, ejection fraction and BNP." *PLoS One* 9.1 (2014): e81699.
9. Many, many more…..

Outline

1. Dichotomizing Predictor
2. Dichotomizing Endpoint
3. Simulations
4. Real data illustration

Dichotomizing a Continuous Predictor

Dichotomizing is Unrealistic Unless Reality is a Step Function



Can Occur in Biology

Not Handled by Dichotomization

Unlikely to Occur

Assumed in Much of Biomarker Research

Figure 18.2: Two kinds of thresholds. The pattern on the left represents a discontinuity in the first derivative (slope) of the function relating a marker to outcome. On the right there is a lowest-order discontinuity.

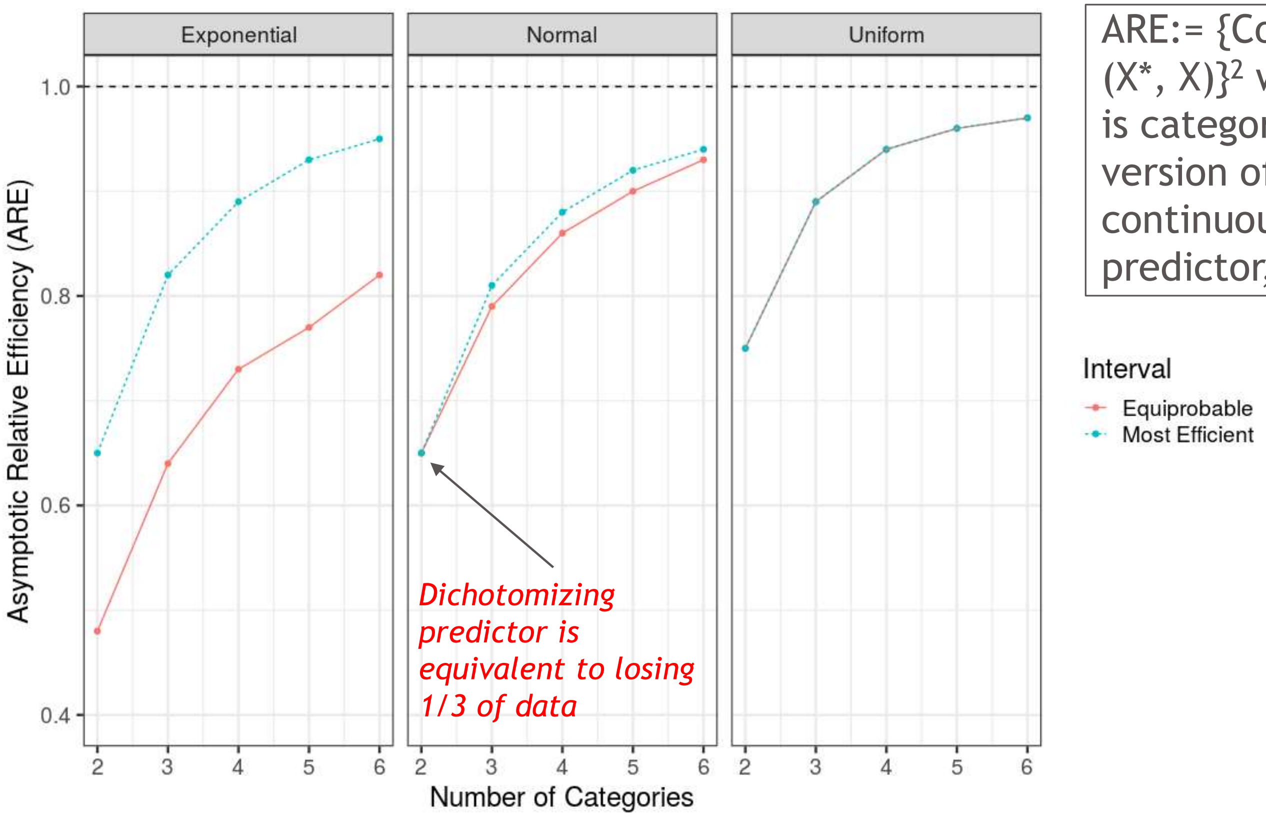
Figure copied from:

http://hbiostat.org/doc/bbr.pdf (Frank Harrell)

Loss of efficiency when categorizing a continuous predictor

3 important points from this figure:

1. Models are less efficient when categorizing continuous predictor
2. If you must categorize (Why?), the more categories the better
3. 3 levels is much better than 2
4. Equiprobable intervals are NOT necessarily optimal



Dichotomizing predictor is equivalent to losing 1/3 of data

ARE:= {Correlation (X*, X)}^{2} where X* is categorized version of continuous predictor, X

Figure based on table from:

Lagakos, S. W. "Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable." *Statistics in medicine* 7.1-2 (1988): 257-274.

Loss of Information Causes Failure to Adjust for Imbalance in Prognostic Factor

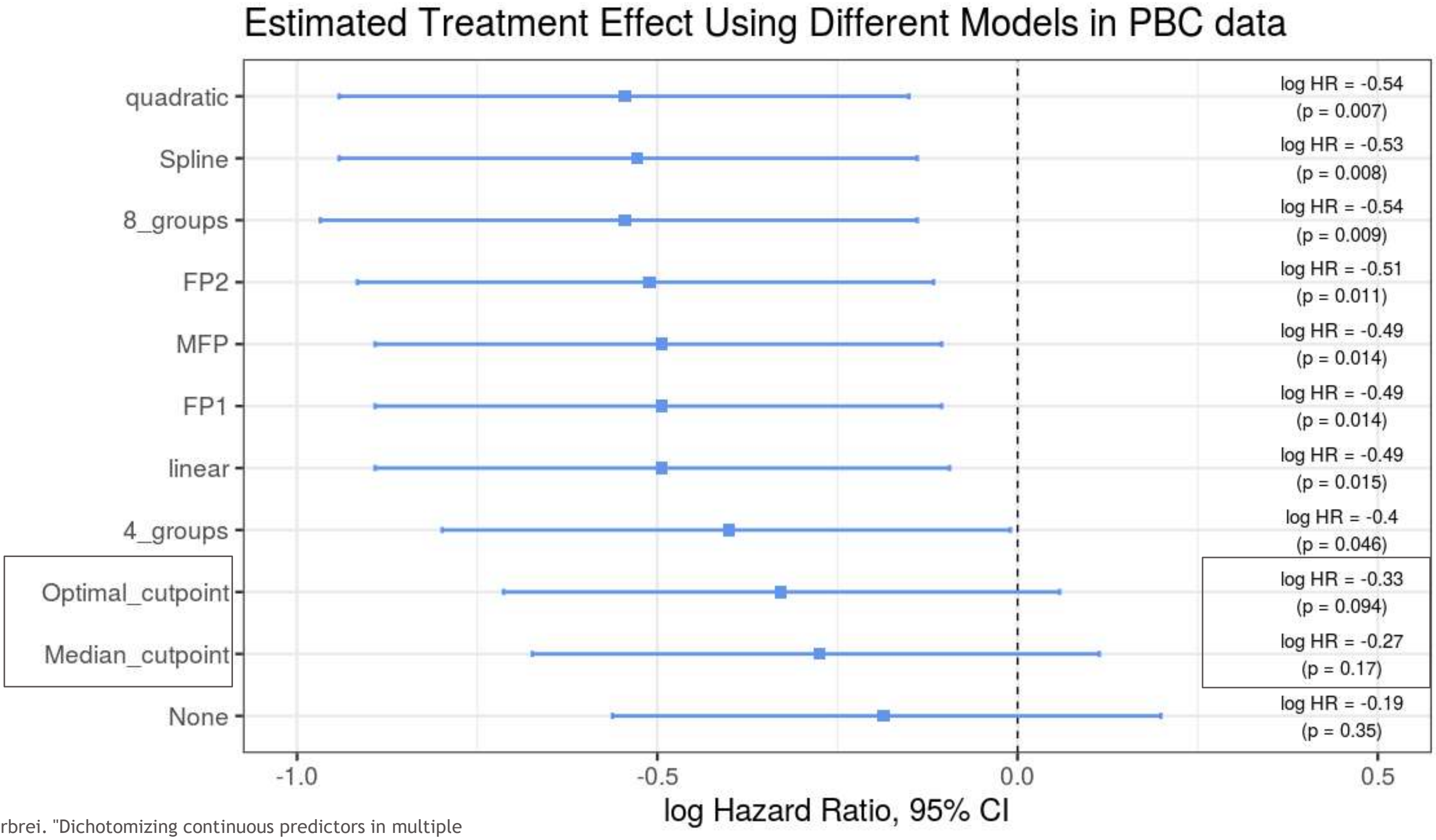
None:= No adjustment for bilirubin

*n*_groups:= Number of categories bilirubin is split into

FP*n*:= Fractional Polynomial of nth degree for bilirubin

MFP:= Multivariable Fractional Ploynomial for bilirubin

Optimal_cutpoint:= Single cutpoint for bilirubin determined based on outcome



Models built containing 3 clinical covariates and various transformations of bilirubin

Royston, Patrick, Douglas G. Altman, and Willi Sauerbrei. "Dichotomizing continuous predictors in multiple regression: a bad idea." *Statistics in medicine* 25.1 (2006): 127-141.

Type I Error Inflation when Categorizing Continuous Confounder

Type I error rate for a test of effect of X1 on logit(p), for different transformations of X2, vs. correlation between X1 and X2

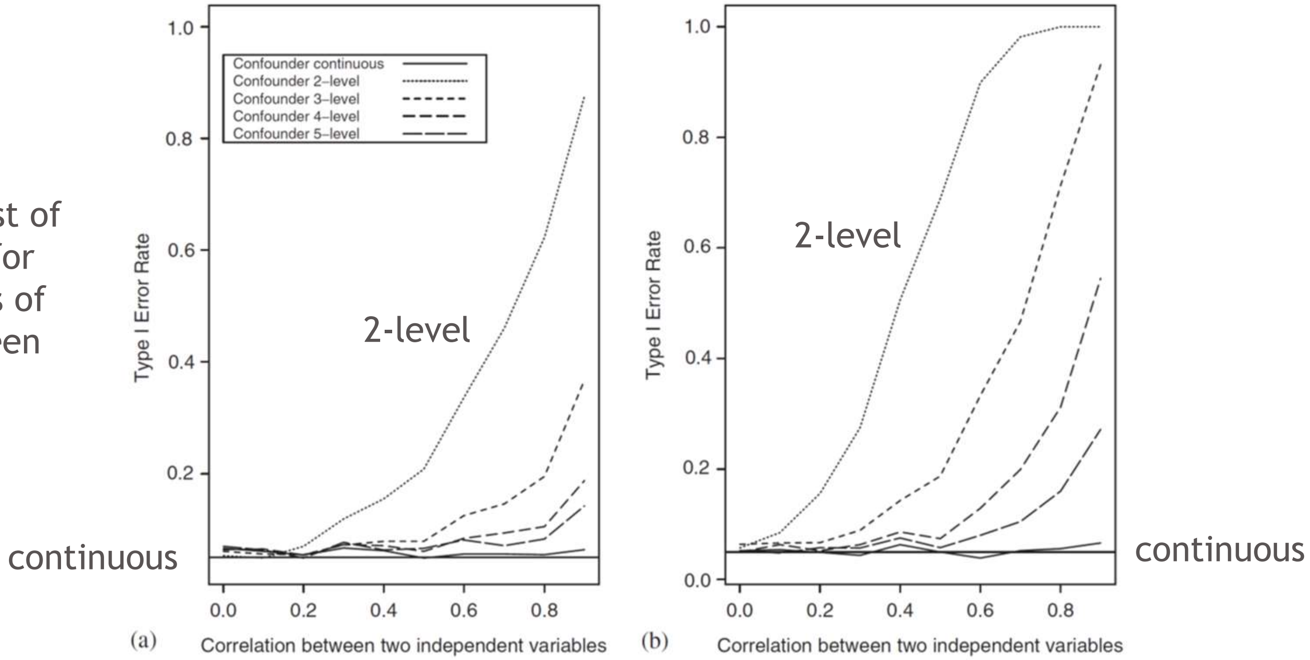


Figure 3. Observed type I error rate (confounder categorized): (a) sample size 100; and (b) sample size 500.

Figure copied from:

Austin, Peter C., and Lawrence J. Brunner. "Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses." *Statistics in medicine* 23.7 (2004): 1159-1178.

Different “Approaches” to Dichotomizing Continuous Predictors

- Using the response variable, “Y” (e.g. OS, BOR, continuous marker)
  - Search for “optimal” cutpoint
- Not using “Y”
  - Use median of distribution or arbitrary “round” number

Disadvantages of Using “Optimal” or “Arbitrary” Cutpoint

- Usually will not generalize well to external data
  - Arbitrary cutpoint typically is based on distribution in available (training) sample
  - Optimal cutpoint is often not “optimal” in external data
  - Optimal cutpoint may have wide confidence interval and therefore may not be clinically meaningful
- Can artificially increase Type I error rate (see Figure on slide 11)
- Can increase Type II error rate as well (see Figure on slide 9)

Dichotomizing Continuous Endpoint

Problems Associated with Dichotomizing Endpoint

- Greater risk of over-fitting due to decreased effective sample size (see below)

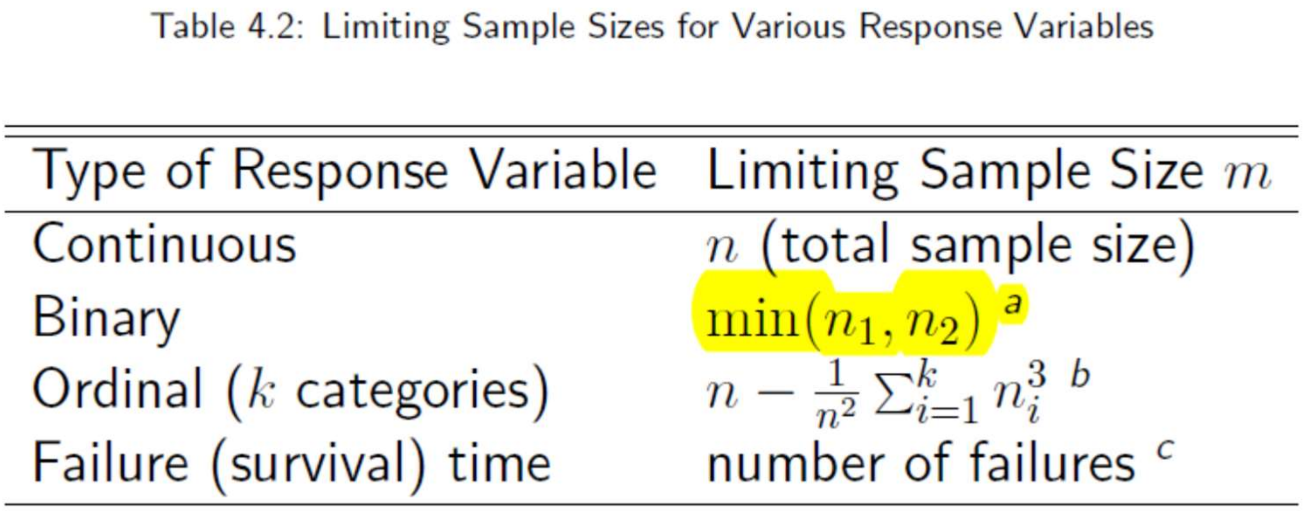


Table 4.2: Limiting Sample Sizes for Various Response Variables

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ [a]
Ordinal (k categories)	$n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$ [b]
Failure (survival) time	number of failures [c]

Table copied from:

Harrell Jr, Frank E. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis.* Springer, 2015.

Dichotomizing Endpoint Wastes Samples

- HUGE loss of information – if median of outcome is chosen <u>*a priori*</u> as the cutpoint for a “normal” variable, there is a 36% loss of information
    - e.g., with n = 100 then n* = 64 after dichotomization

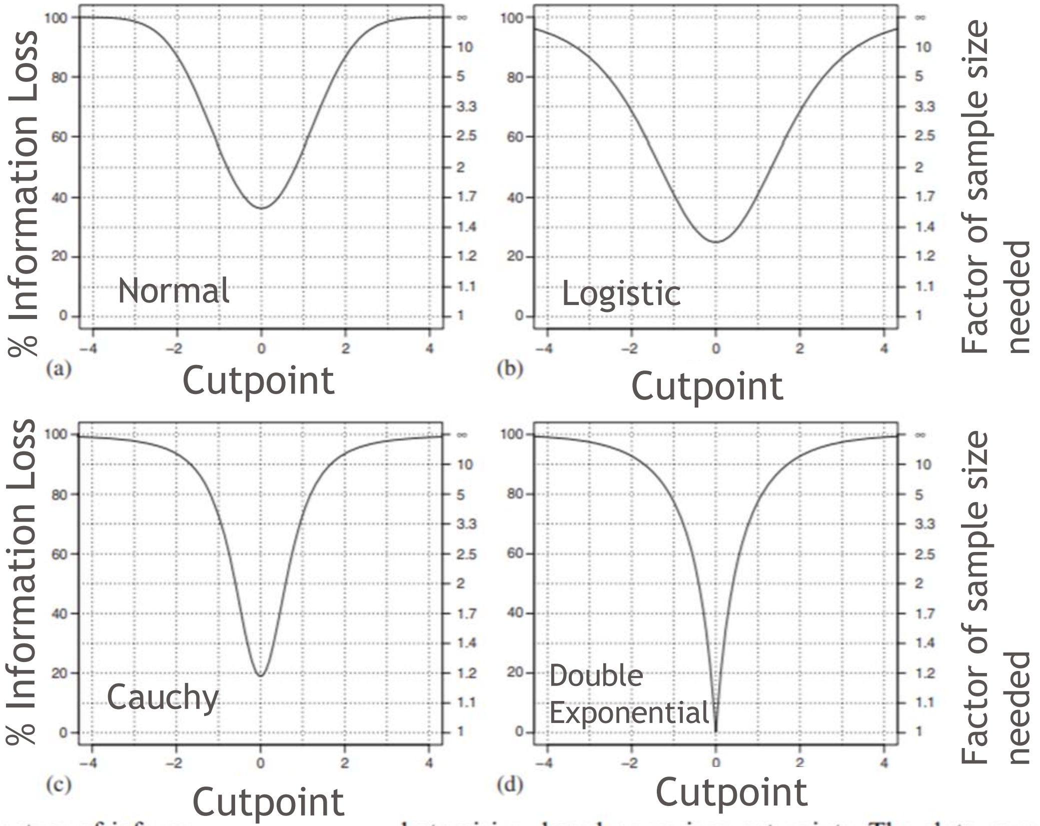


Figure 1. Percentage of information lost when dichotomizing, based on various cut points. The plots correspond to normal (a), logistic (b), Cauchy (c), and double exponential (d) distributions. The right vertical axes correspond to the factor of sample size increase needed to mitigate the loss.

Figure copied from:

Fedorov, Valerii, Frank Mannino, and Rongmei Zhang. "Consequences of dichotomization." *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry* 8.1 (2009): 50-61.

Dichotomania Simulations: Regression Splines to the Rescue

- Simulated 1000 replicate datasets, each with N = 500 subjects
- Simulated High and Low Signal to Noise Ratios (SNR) for each of 3 distinct forms of relationship between a continuous predictor (x) and continuous outcome (y), resulting in 6 total scenarios:
  1. True Relationship is Step Function
     A. High SNR
     B. Low SNR
  2. True Relationship is Linear Function
     A. High SNR
     B. Low SNR
  3. True Relationship is Linear Spline Function
     A. High SNR
     B. Low SNR

Step Function: High Signal to Noise Ratio

Continuous Test	Dichotomized Test	Noise Test
y ~ continuous_x + dichotomized_x	y ~ continuous_x + dichotomized_x	y ~ continuous_x + noise
y ~ dichotomized_x	y ~ continuous_x	y ~ continuous_x

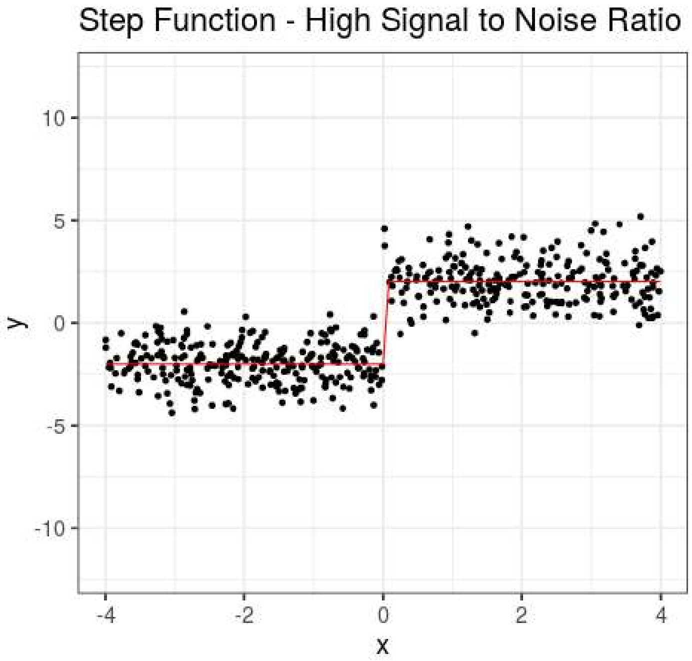
Beta = -2, 2

Sigma = 1

True Step: x = 0

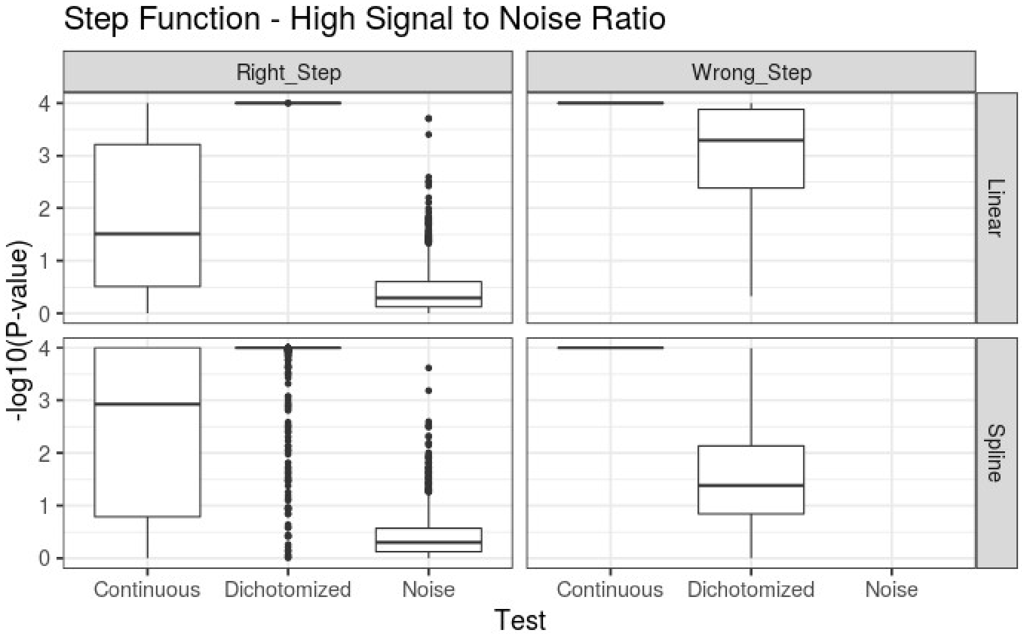
Wrong Step: x = 2

Truth: y ~ Beta1*I(x<=0) + Beta2*I(x>0) + N(0, Sigma)



n = 500

Likelihood Ratio Test



n = 500
Simulations = 1000

Step Function: Low Signal to Noise Ratio

Continuous Test	Dichotomized Test	Noise Test
y ~ continuous_x + dichotomized_x	y ~ continuous_x + dichotomized_x	y ~ continuous_x + noise
y ~ dichotomized_x	y ~ continuous_x	y ~ continuous_x

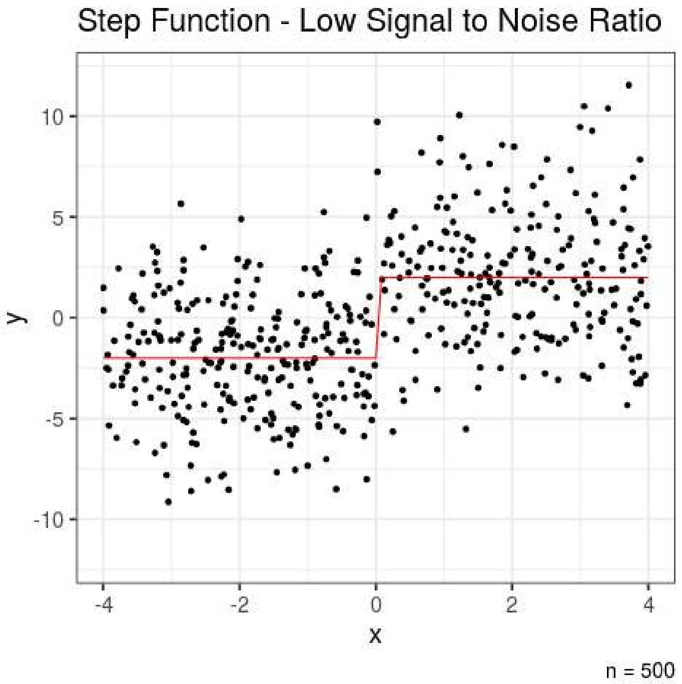
Beta = -2, 2

Sigma = 3

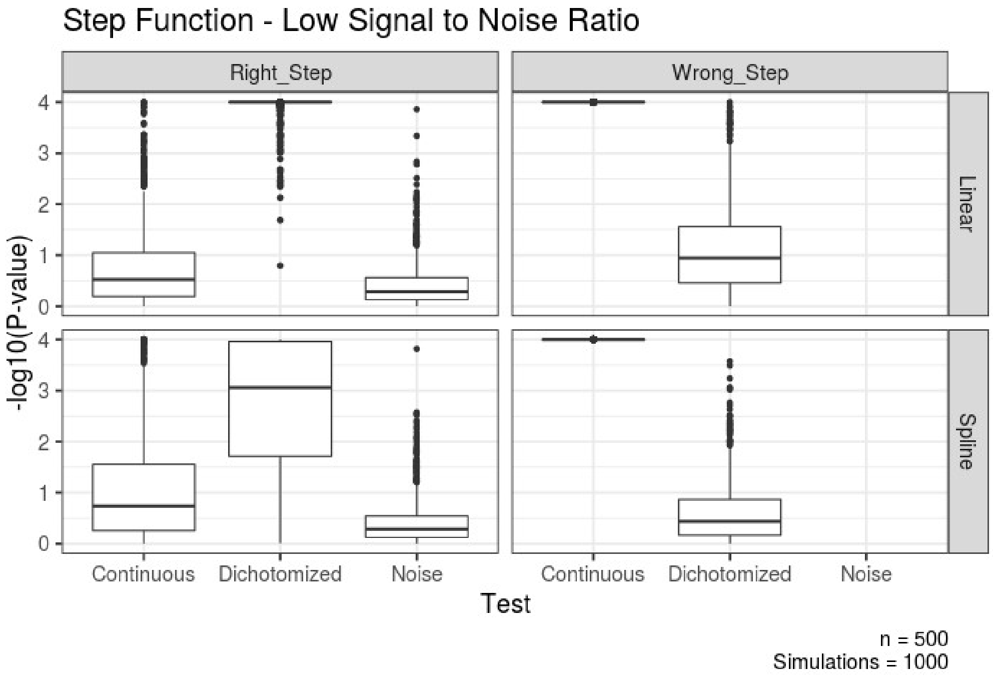
True Step: x = 0

Wrong Step: x = 2

Truth: y ~ Beta1*I(x<=0) + Beta2*I(x>0) + N(0, Sigma)



Likelihood Ratio Test



Linear Function: High Signal to Noise Ratio

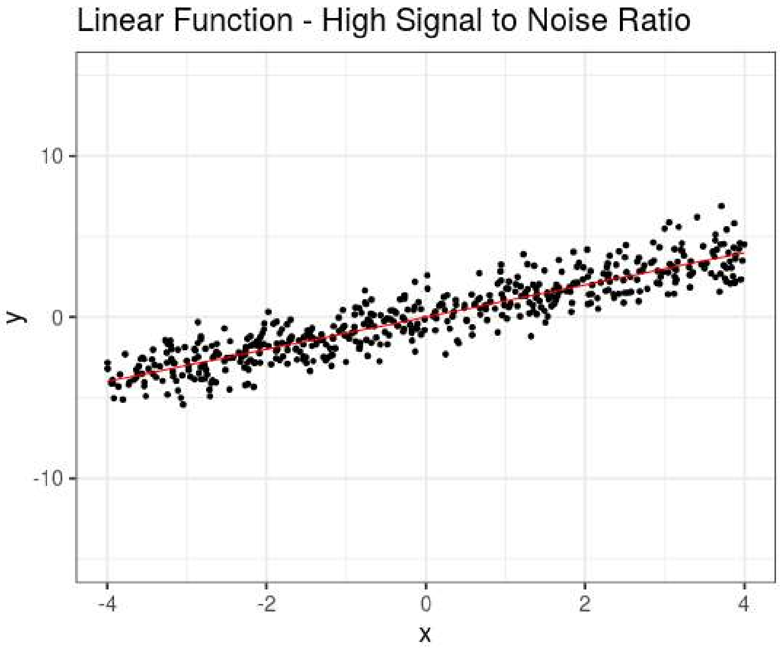
Continuous Test	Dichotomized Test	Noise Test
y ~ continuous_x + dichotomized_x	y ~ continuous_x + dichotomized_x	y ~ continuous_x + noise
y ~ dichotomized_x	y ~ continuous_x	y ~ continuous_x

Beta = 1

Sigma = 1

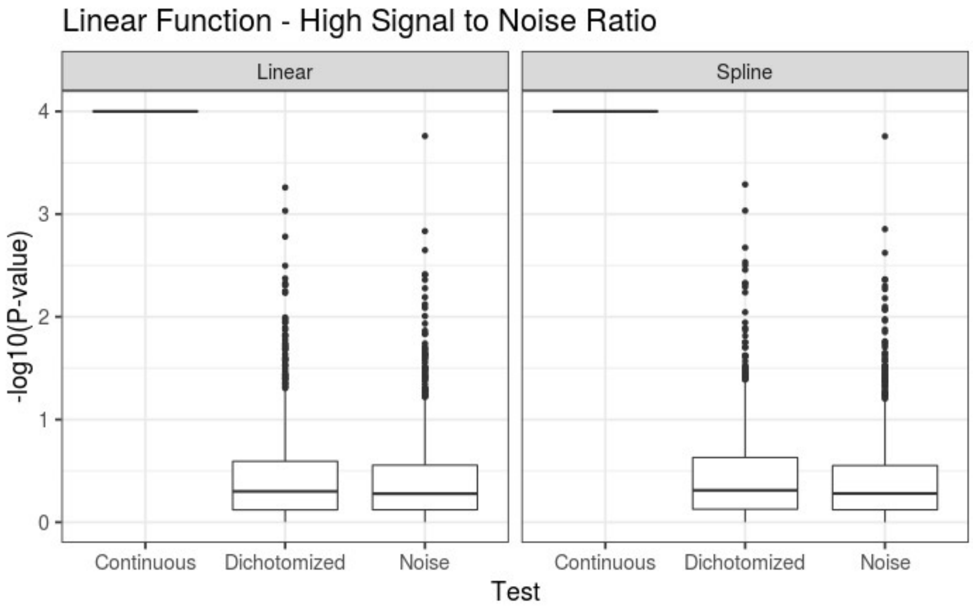
Cutoff: x = 0

Truth: y ~ Beta*x + N(0, Sigma)



n = 500

Likelihood Ratio Test



n = 500
Simulations = 1000

Linear Function: Low Signal to Noise Ratio

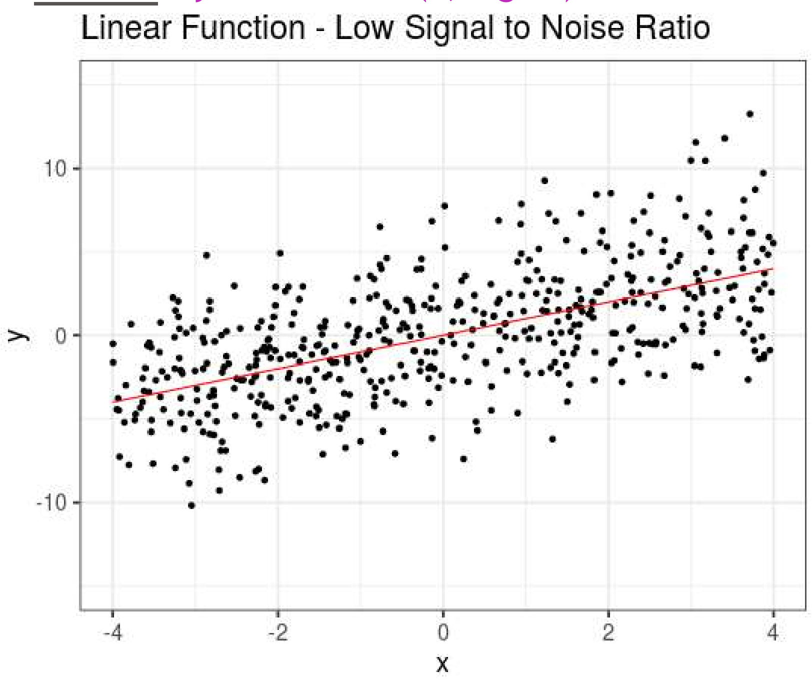
Continuous Test	Dichotomized Test	Noise Test
y ~ continuous_x + dichotomized_x	y ~ continuous_x + dichotomized_x	y ~ continuous_x + noise
y ~ dichotomized_x	y ~ continuous_x	y ~ continuous_x

Beta = 1

Sigma = 3

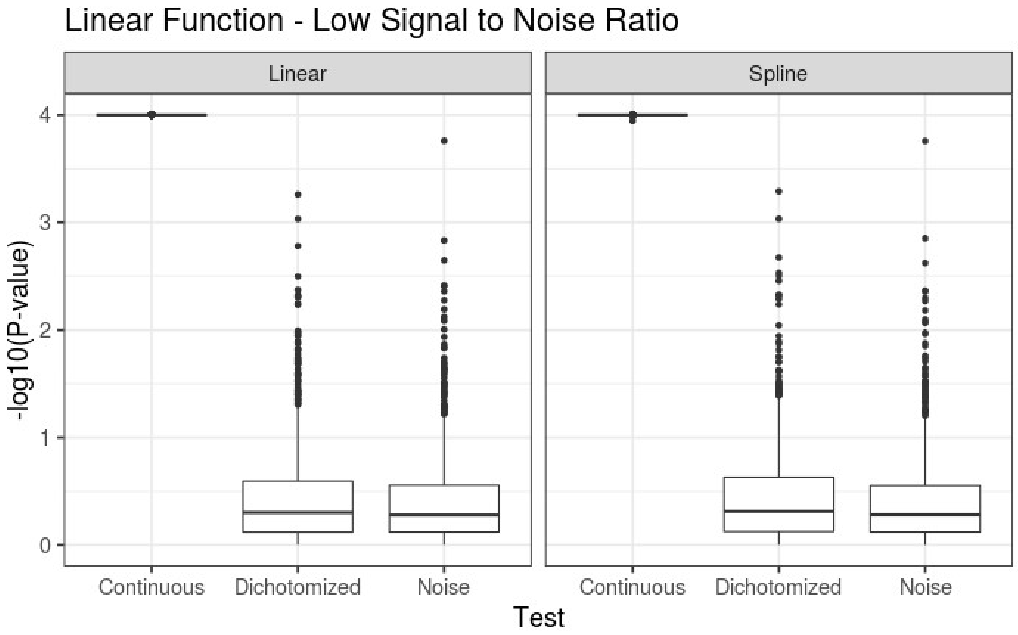
Cutoff: x = 0

Truth: y ~ Beta*x + N(0, Sigma)



n = 500

Likelihood Ratio Test



n = 500
Simulations = 1000

Linear Spline Function: High Signal to Noise Ratio

Continuous Test	Dichotomized Test	Noise Test
y ~ continuous_x + dichotomized_x	y ~ continuous_x + dichotomized_x	y ~ continuous_x + noise
y ~ dichotomized_x	y ~ continuous_x	y ~ continuous_x

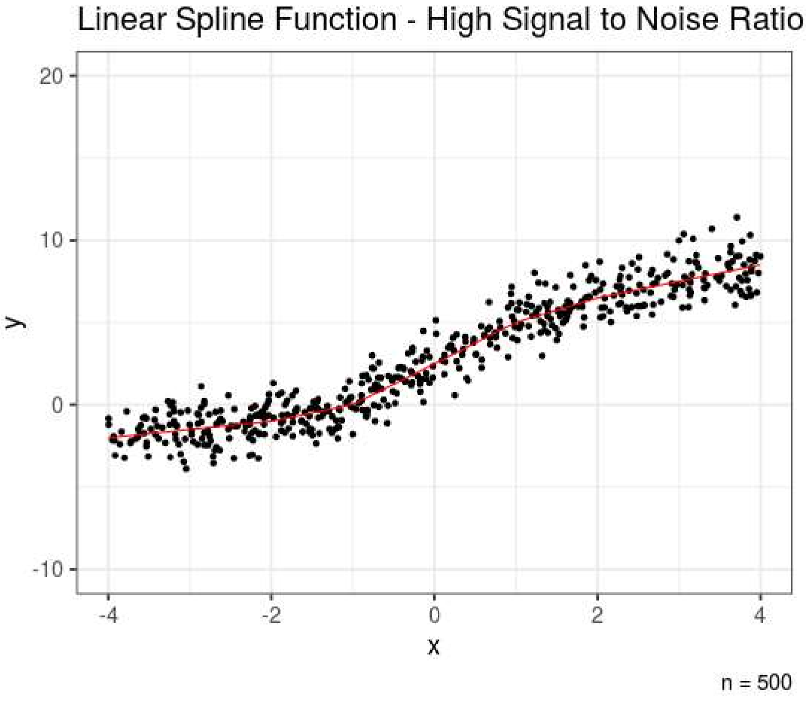
Knots: -2, -1, 1, 2

Beta: 0.5, 0.5, 1.5, -1, -0.5

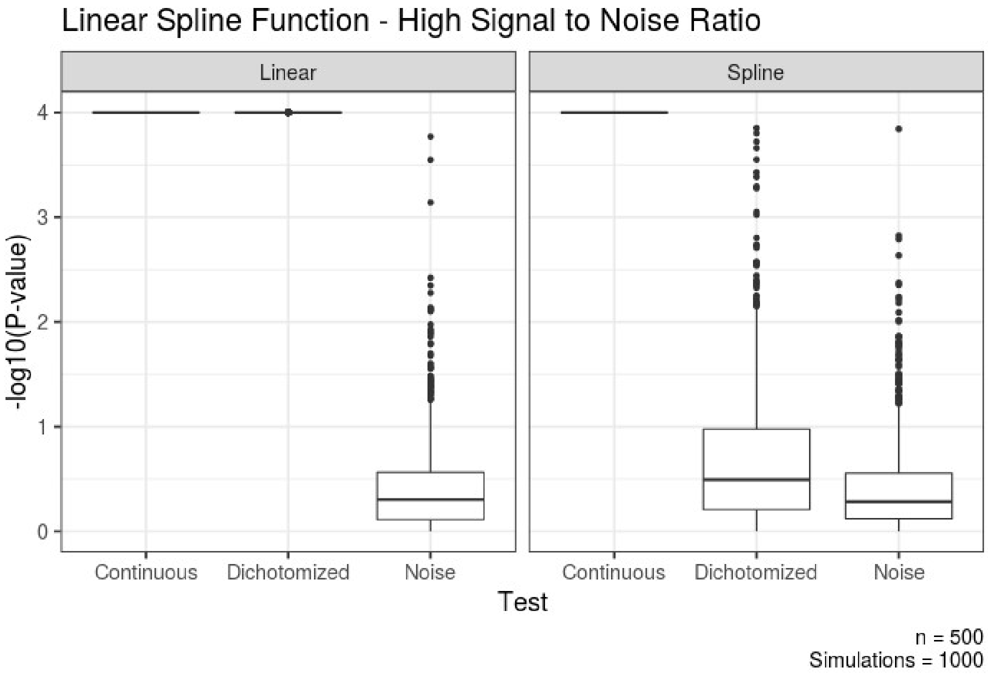
Sigma = 1

Cutoff: x = 0

Truth



Likelihood Ratio Test



Linear Spline Function: Low Signal to Noise Ratio

Continuous Test	Dichotomized Test	Noise Test
y ~ continuous_x + dichotomized_x	y ~ continuous_x + dichotomized_x	y ~ continuous_x + noise
y ~ dichotomized_x	y ~ continuous_x	y ~ continuous_x

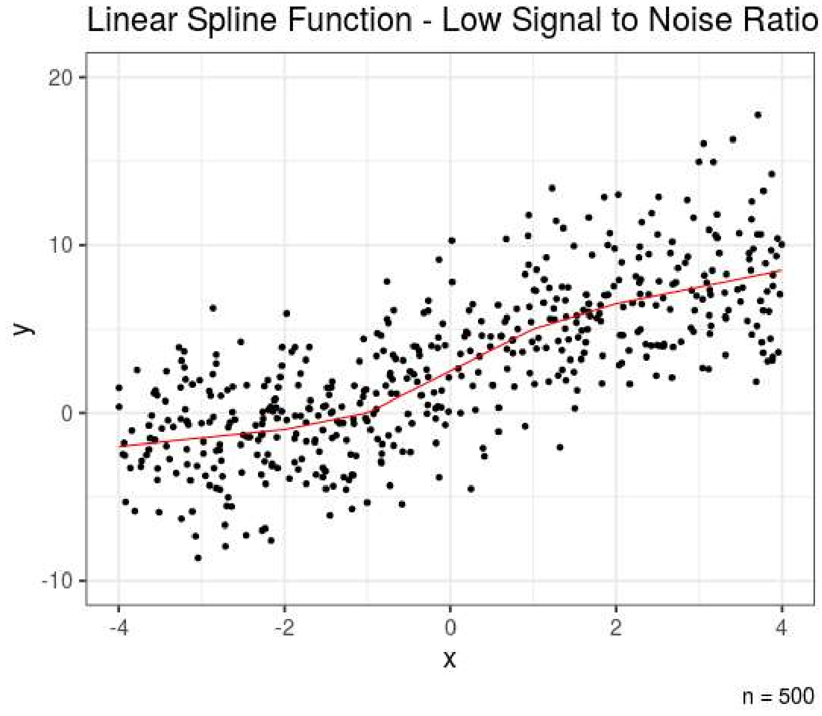
Knots: -2, -1, 1, 2

Beta: 0.5, 0.5, 1.5, -1, -0.5

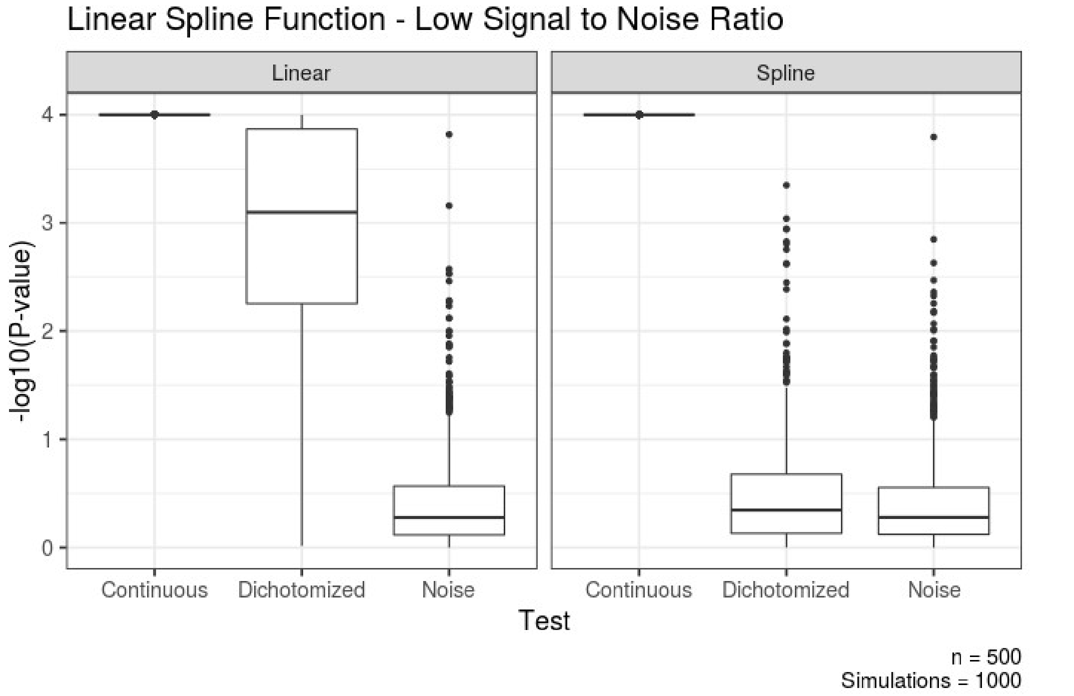
Sigma = 3

Cutoff: x = 0

Truth



Likelihood Ratio Test



Common Reasons Researchers Tend to Dichotomize Continuous Predictors - *And the Truth Behind Them*

Reason	Truth
Simplicity	*Dichotomization actually makes it more difficult to generalize and interpret results due to different cutpoints in different datasets.* *Visualizations can also be more difficult to process (see slide 27).*
Don’t want to assume linear relationship between predictor and outcome	*Step function is less plausible than linearity assumption (see slide 8).* *Restricted Cubic Splines provide a better alternative to linearity assumption than dichotomizing.*
Anticipate eventual need to provide cutoff for predictor	*“Optimal” cutoff usually will not generalize to other data. Better to dichotomize, if necessary, the linear predictor score (i.e., $\hat{Y}$) instead of individual predictors.*

Summary

- Dichotomizing can obscure important patterns in data
- Dichotomous patterns usually are unrealistic
- Models with categorized continuous predictors are less efficient (waste information)
- Loss of information by categorizing can cause failure to adjust for imbalance in prognostic factors
- Categorizing continuous confounders can lead to inflation of Type I error rates
- Using an “optimal” or arbitrary cut point
  - Usually, will not generalize well to external data
  - Can inflate both Type I and Type II error rates
  - An “optimal” cut point may have large uncertainty (e.g., wide confidence interval)
- Dichotomizing a continuous endpoint reduces effective sample size (i.e., wastes samples)
  - This can increase risk of over-fitting and decrease power
- Common excuses for dichotomizing are based on fallacies