THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

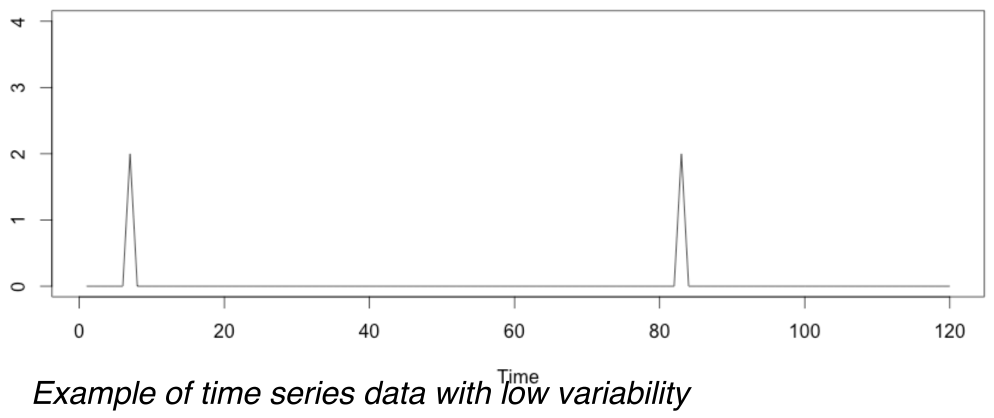
Detecting low variability

Cara Arizmendi and Kathleen Gates, PhD.

University of North Carolina at Chapel Hill • L.L. Thurstone Psychometric Laboratory

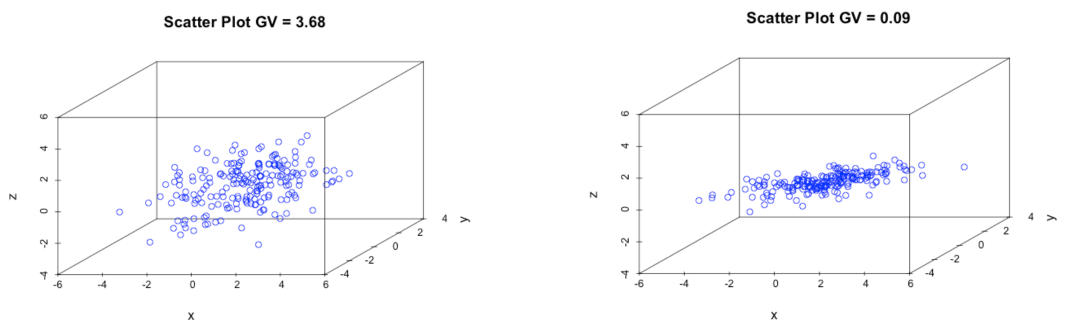
Introduction

- Low variability in daily diary data is a common problem.
- The method for deciding whether to remove a low-variance variable is often unreported (e.g., Admon, 2013; Kendler, 2017). Sometimes researchers will report removing variables with zero or "near zero" variance (Whitley, 2000; Zevon,1982). A standard that address more specifically, low variability, would be useful.



Example of time series data with low variability

- Assessing variance on a univariate basis does not account for dependencies between variables in multivariate analyses. Low variance can impact estimates and inferences due to inflation in precision matrix (Kutner, 2005).
- Generalized variance (GV), or the determinant of the covariance matrix was first proposed by Wilks (1932) as a measure of multivariate scatter. A GV of zero indicates at least one variable has no variance. A GV approaching zero indicates at least one variable has variance approaching zero.



- Haitovsky's test for multicollinearity (1969), similarly, uses the determinant of the correlation matrix as a measure of collinearity. However, the GV is not standardized like the Haitovsky statistic. Additionally, the distribution is unknown in many situations (e.g., large number of variables, non-normal data) making the statistic difficult for application of a test.
- Using bootstrapping may be useful as a test of the GV (Sengupta, 2011). However, it is necessary to determine if a dataset with GV deemed too small is actually problematic.

Objectives

1. Can bootstrapping allow us to test whether the GV is too close to zero?
2. Do the results we get from bootstrapping show that low variability is a problem in linear regression? In other words, can including a low-varying variable in a model inflate the variability of estimates enough to lead us to incorrect inferences?

Methods

Objective 1: Bootstrapping the GV

- Multivariate normal data (n=80) was generated with three variables, then rounded to the nearest integer to simulate Likert-type data. One variable was constant with μ = 0,.
- For the constant variable, observations were randomly selected to vary from the constant of zero with a value of 1.
- Observations were manipulated to vary by 2, 3, 4, and 5 observations, for a total of four datasets to be bootstrapped. Covariance matrices for each condition are below:

$$\text{nobs varying} = 2 \quad \begin{bmatrix} 1.96 & 0.44 & -0.01 \\ 0.44 & 4.06 & -0.03 \\ -0.01 & -0.03 & 0.02 \end{bmatrix}$$

$$\text{nobs varying} = 3 \quad \begin{bmatrix} 1.81 & 0.49 & 0.06 \\ 0.49 & 4.20 & -0.04 \\ 0.06 & -0.04 & 0.04 \end{bmatrix}$$

$$\text{nobs varying} = 4 \quad \begin{bmatrix} 2.10 & 0.54 & -0.02 \\ 0.54 & 4.23 & 0.00 \\ -0.02 & 0.00 & 0.05 \end{bmatrix}$$

$$\text{nobs varying} = 5 \quad \begin{bmatrix} 2.08 & 0.56 & 0.01 \\ 0.56 & 4.11 & -0.09 \\ 0.01 & -0.09 & 0.06 \end{bmatrix}$$

- Bootstrapping was conducted as described in Sengupta (2011).
- Confidence intervals were calculated around the true GV.

Objective 2: Do results from Objective 1 indicate that low variability is problematic?

- The true covariance matrices from Objective 1 were used to generate data. Sample size was also varied at n= 30, 60, and 100, for a total of 12 conditions.
- This process was replicated 1000 times per condition.
- Each data set was fit to two linear models, where y is a normal variance variable, X_1 is a normal variance variable, and X_2 is a low variance variable:

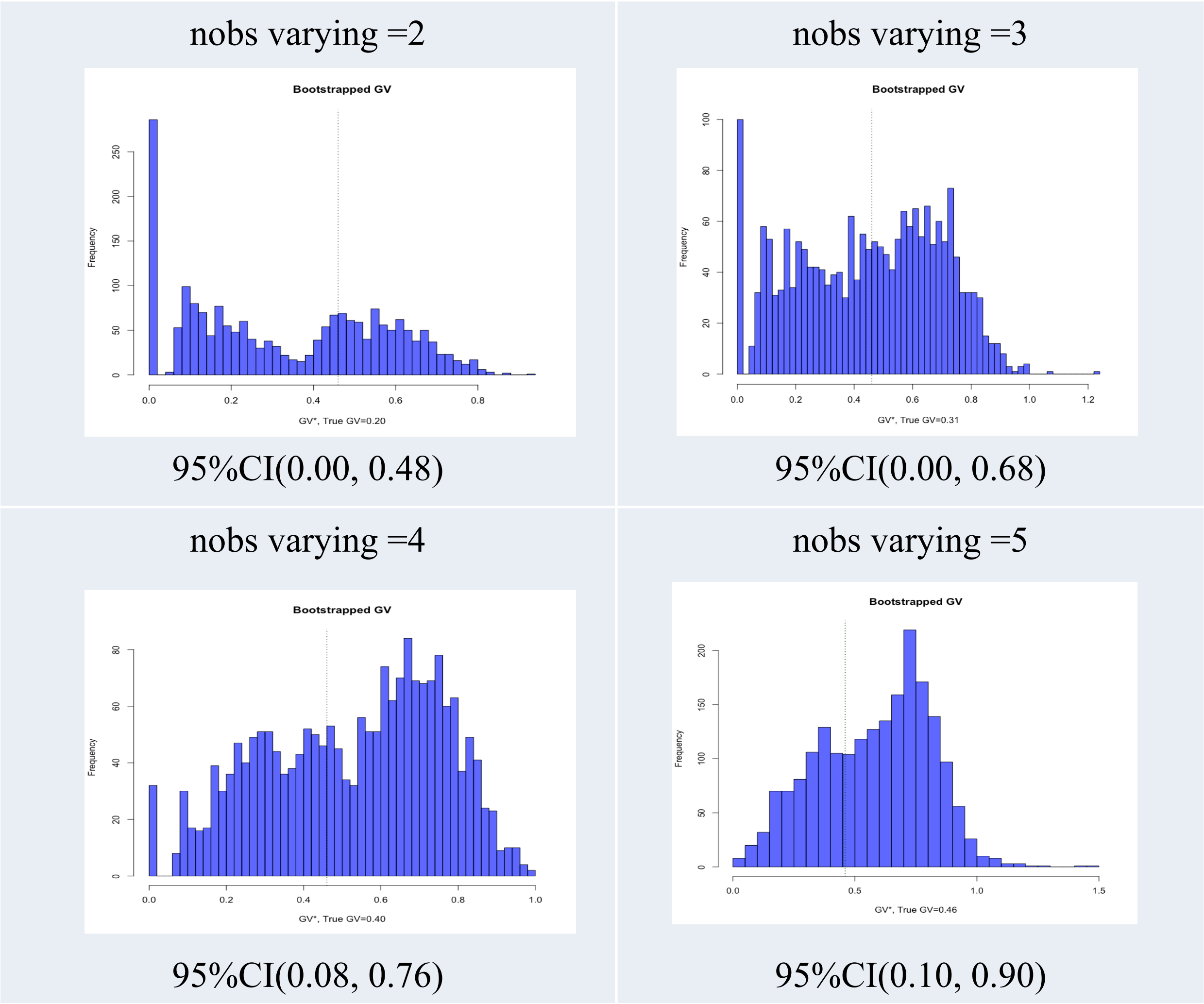
$$\text{Model 1. } y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

$$\text{Model 2. } y = \beta_0 + \beta_1 X_1$$

- The t-statistic for X_1 was recorded for both models to determine if large discrepancies exist in inferential estimates when including a low variance variable or not including a low variance variable in the model.

Results

Objective 1: Bootstrapping the GV



95%CI(0.00, 0.48)

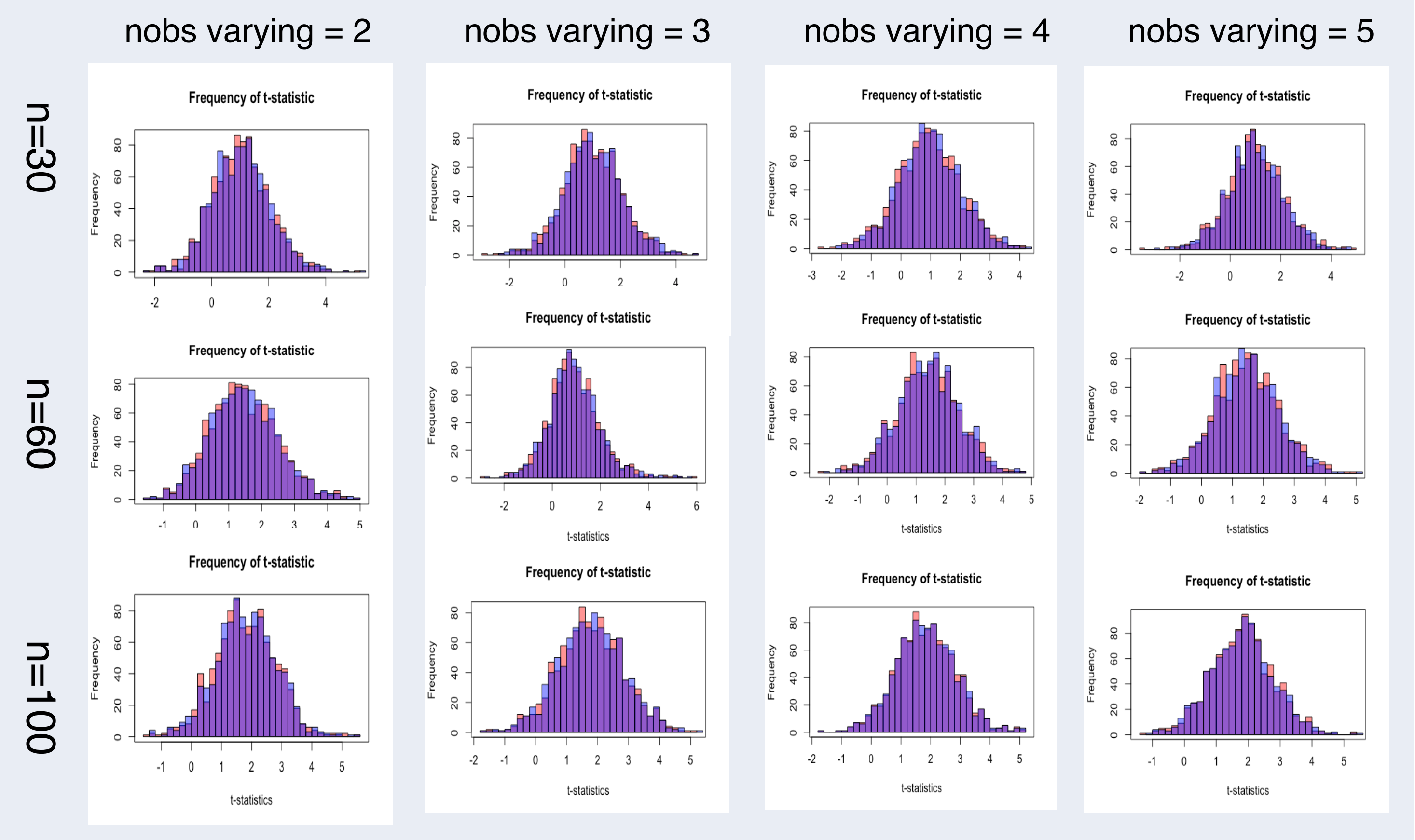
95%CI(0.00, 0.68)

95%CI(0.08, 0.76)

95%CI(0.10, 0.90)

- Based on testing the null hypothesis that the GV is zero, we fail to reject the null that the GV is zero when only 2 or 3 observations are varying from an otherwise constant variable. Based on these results, we need at least 4 observations varying from a constant variable to obtain a large enough GV.
- Separate tests suggest these findings hold at varying sample sizes.

Objective 2: Do results from Objective 1 indicate that low variability is problematic?



- The true beta weight for X_1 was 0.10.
- The figure above displays histograms of the t-statistics obtained in each condition over 1000 iterations. Blue indicates t-statistics obtained when Model 2 is used. Red indicates t-statistics obtained in Model 2. Purple indicates overlap between the two models.
- Overall, there is a lot of overlap between t-statistics obtained with Model 1 and Model 2, regardless of sample size or number of observations varying from the constant of zero in $X_{2.}$
- Deviations between the models are few and do not demonstrate a systematic pattern.

Conclusions

- Using bootstrapping to test whether a GV is too low may not be sufficient for detecting low variability.
- When we have 4 or more observations varying from the constant, we find confidence intervals that do not include zero, regardless of sample size. We may need to explore subsetting to see if the confidence intervals will vary more according to sample size.
- Exploratory analyses indicate that low variability may not have a deleterious impact on inferential statistics. More simulations are needed to determine the limits of these findings.

References

- Admon, R., Leykin, D., Lubin, G., Engert, V., Andrews, J., Pruessner, J., & Hendler, T. (2013). Stress-induced reduction in hippocampal volume and connectivity with the ventromedial prefrontal cortex are related to maladaptive responses to stressful military service. *Human Brain Mapping*, *34*(11), 2808–2816.
- Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics*, *51*(4), 486–489.
- Kendler, K. S., & Aggen, S. H. (2017). Symptoms of major depression: Their stability, familiality, and prediction by genetic, temperamental, and childhood environmental risk factors. *Depression and Anxiety*, *34*(2), 171–177.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw Hill.
- SenGupta, A., & Ng, H. K. T. (2011). Nonparametric test for the homogeneity of the overall variability. *Journal of Applied Statistics*, *38*(9), 1751–1768.
- Whitley, D. C., Ford, M. G., & Livingstone, D. J. (2000). Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *Journal of Chemical Information and Computer Sciences*, *40*(5), 1160–1168.
- Wilks, S. S. (1932). Certain Generalizations in the Analysis of Variance. *Biometrika*, *24*(3/4), 471–494.
- Zevon, M. A., & Tellegen, A. (1982). The Structure of Mood Change: An Idiographic/Nomothetic Analysis. *Journal of Personality and Social Psychology*, *43*(1), 111–122.