DIAGNOSTICS FOR DETECTING LOW VARIABILITY

Cara Arizmendi

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Psychology and Neuroscience.

Chapel Hill
2018

Approved by:

Kathleen Gates

David Thissen

Kenneth Bollen

ABSTRACT

CARA ARIZMENDI: Diagnostics for detecting low variability.
(Under the direction of Kathleen Gates)

Whether to include variables with low variability in analyses is a predicament for researchers. Like the case of multicollinearity in which a correlation matrix has a near-zero determinant, data with low-variability variables will have a near-zero determinant of the covariance matrix. Similar to problems with multicollinearity, we may expect that retaining low-variability variables in our models would result in unstable estimates of the other variables in the model. Despite this predicament: 1. a method for identifying these variables for removal has yet to be developed and 2. the impact of low variability on estimation of other variables has yet to be explored. We present a simulation study exploring the impact of low variability in linear regression, while also assessing indices of low variability. Simulations suggest the stability of estimates of normal-variability variables, despite the presence of a low variability variable and the instability of estimates for variables with low variability.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

Obtaining data where one or more variables has low variability is a common problem for researchers in the social and behavioral sciences, especially when collecting daily diary data. Whether survey data is collected in which most respondents select the same answer, a participant's mood rarely varies across the course of a daily diary study, or a period of movement data is collected at a time when the participant is mostly sedentary, researchers are likely to encounter low-variability variables. When encountering a low-variability variable, the question arises of whether the variable should be included or excluded from analyses. While creating more reliable, sensitive scales may in some cases alleviate this problem, here we explore indices for deciding when to exclude low-variability variables and the impact of low-variability variables on estimation when the data have already been collected. We first discuss the role of low variability in estimating regression coefficients followed by a description of the characteristics of covariance matrices where one or more variables has low variability. We will then provide a demonstration of the proposed indices for detecting low variability, generalized variance (GV) (Wilks 1932) which was motivated by the Haitovsky test for high multicollinearity (Haitovsky 1969) and the Gini coefficient (Gini 1921). Next, we will present a simulation demonstrating the performance of each index and the impact of low variability on standard errors. Finally, we conclude with an empirical example using daily diary data, varying the number of variables included in analysis.

1.1 Impact of low variance in regression

First, we should distinguish between low variability and low variance. The focus of this paper is low variability; however, a variable with low variability inherently has low variance. A solution for dealing with low variance may be to rescale the variable. Rescaling, however, is not useful when a variable has low variability. In this section, we focus on the impact of low variance in general, but our interest is more specifically in dealing with low variability. We will also refer to low variance in later sections where the discussion is generalizable to both low variance and low variability.

Depending on the severity, it is possible that low variance could present major issues for regression coefficient estimates and the conclusions we draw from those estimates when doing multiple regression. Just as the correlation matrix has a determinant of zero in the case of perfect multicollinearity, the covariance matrix has a determinant of zero in the case of one variable having a variance of zero. Thus, it may follow that the effect of high multicollinearity on the standard errors of estimates is similar to the effect of low variance on the standard errors of estimates. It can be a difficult decision to exclude variables of interest from analyses, especially ones that are theoretically meaningful; however, there may be cases where the impact of low variance supports removal of the variable, specifically, when a low-variance variable impacts inferences. To understand the impact of low variance, we demonstrate how low variance might influence the standard errors of regression coefficients, or β, in linear regression.

First, we begin with a 2X2 covariance matrix consisting of predictor variables *i* and *j*, where on-diagonal elements are the variance of each variable and off-diagonal elements are the covariance between variables:

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{ii}^2 & \sigma_{ij}^2 \\ \sigma_{ji}^2 & \sigma_{jj}^2 \end{bmatrix} \tag{1.1}$$

A key part of many analytic techniques is the inverse of $\mathbf{\Sigma}$, also known as the precision matrix (Fox & Monette 1992; Stevens 1998). We obtain the precision matrix by multiplying the inverse of the determinant of $\mathbf{\Sigma}$ by the adjugate of $\mathbf{\Sigma}$, shown in the following two-variable example:

$$\mathbf{\Sigma}^{-1} = \frac{1}{\det \mathbf{\Sigma}} \begin{bmatrix} \sigma_{jj}^2 & -\sigma_{ij}^2 \\ -\sigma_{ji}^2 & \sigma_{ii}^2 \end{bmatrix} = \frac{1}{\sigma_{ii}^2\sigma_{jj}^2 - \sigma_{ij}^2\sigma_{ji}^2} \begin{bmatrix} \sigma_{jj}^2 & -\sigma_{ij}^2 \\ -\sigma_{ji}^2 & \sigma_{ii}^2 \end{bmatrix} \tag{1.2}$$

1.1.1 Precision matrix under constant variance

When one predictor variable has zero variance, $\mathbf{\Sigma}$ is not invertible. We can immediately see that in covariance matrix $\mathbf{A}$ where $\sigma_j^2 = 0$, the matrix is not positive definite, meaning it cannot be inverted. This is due to the matrix having a determinant of zero. It is impossible to invert a matrix with a determinant of zero, as dividing by zero is not a valid expression.

$$\det \mathbf{A} = \begin{vmatrix} 4 & 0 \\ 0 & 0 \end{vmatrix} = 4 * 0 - 0 * 0 = 0$$

$$\mathbf{A}^{-1} = \frac{1}{0} \begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix}$$

When a covariance matrix has at least one variable with $\sigma^2 = 0$, the decision of whether to remove a variable is simple. With a variance of zero, we cannot obtain a precision matrix. Therefore, the variable has to be removed from the regression model. However, even if we can invert a covariance matrix and obtain a precision matrix when the variance of one or more variables is very close to zero, it is unknown if we will obtain reliable estimates. The question of what level of variance may contribute to unreliable estimates still remains.

1.1.2 Precision matrix under low variance

Because a covariance matrix that contains a variable with zero variance yields a determinant of zero, it follows that as the variance of at least one variable approaches zero, the determinant will also approach zero. For example, in matrix $\mathbf{B}$ where variable *j* has low variance, the determinant is not zero but is very close to zero.

$$\det \mathbf{B} = \begin{vmatrix} 4 & 0.05 \\ 0.05 & 0.02 \end{vmatrix} = 4 * 0.02 - 0.05 * 0.05 = 0.07$$

Even though we can invert **B**, the low variance yields a small determinant. Ultimately, the small determinant results in a precision matrix with inflated values.

$$\mathbf{B}^{-1} = \frac{1}{0.07}\begin{bmatrix} 0.02 & -0.05 \\ -0.05 & 4 \end{bmatrix} = \begin{bmatrix} 0.26 & -0.65 \\ -0.65 & 51.61 \end{bmatrix}$$

It is helpful to consider how the precision matrix relates to regression analysis. The precision matrix provides the partial correlations on the off-diagonal elements and the precision for the regression coefficients, β, on the on-diagonal elements, often interpreted as the variability of β (Fox & Monette 1992; Stevens 1998). Within the context of high multicollinearity, having highly inflated precision values results in unstable standard errors of β, possibly leading to incorrect inferences about the results of the analyses. Because the covariance matrix under low variance functions similarly to the correlation matrix under high multicollinearity, identifying when a variable's variance is too low in the context of the entire regression model may be important for effective analysis and interpretation (Kutner, Nachtsheim, Neter, & Li 2005).

1.2 Identifying low variability

Researchers often do not report their methods for removing variables with low variability (or variance), only mentioning that a variable was not included in their analyses or was re-operationalized due to the variability (or variance) being too low (e.g., Admon et al. (2013); Kendler & Aggen (2017); Stice, Shaw, & Marti (2007); Gibbons & Roth (2002); Chambers (1985)), so the literature on how to remove low-variability variables is limited. The most common method is to look at the variance of individual variables and remove those that have a variance of zero or close to zero (Whitley, Ford, & Livingstone 2000; Lebo & Nesselroade 1978; Zevon & Tellegen 1982). While this method is practical and quick, there are not clear cutoff points for determining how low the variance has to be for it to be too low. Additionally, this method does not address whether the variable is theoretically meaningful or even explanatory despite having low variance. It is possible that whether a variable's low variance will impact the regression model as a whole is dependent on its relationship to other variables in the model. Using the covariance matrix to make decisions about whether to exclude a given variable allows us to account for shared variance in the model. More specifically, we can use the determinant of $\mathbf{\Sigma}$ to assess for low variance. Here, we look at two ways to assess low variability, the generalized variance (GV) (Wilks 1932) and the Gini coefficient (Gini 1921).

1.2.1 Generalized variance

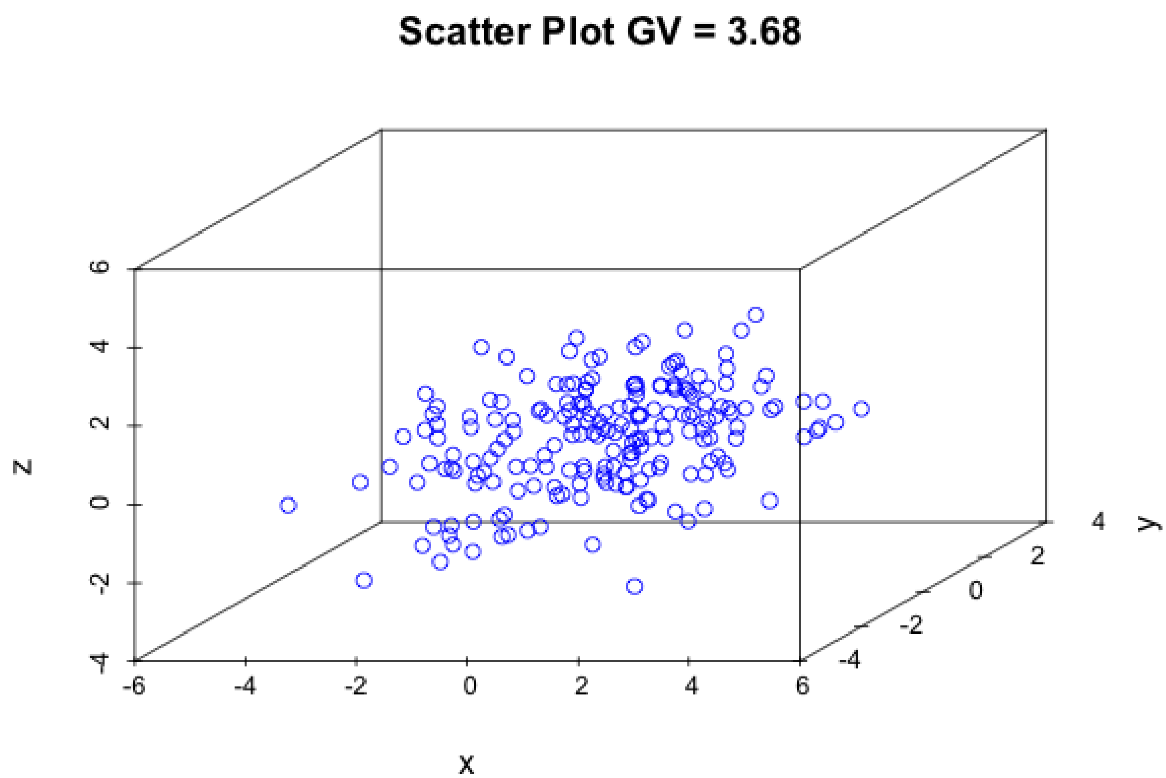
In our discussion of the impact of low variance, we showed how zero variance in at least one variable yields a determinant of zero and low variance in at least one variable yields a determinant approaching zero. There is one other instance where $\mathbf{\Sigma}$ will yield a determinant of zero or close to zero. When the data has perfect multicollinearity, $\mathbf{\Sigma}$ yields a determinant of zero, and when the data have certain cases of very high multicollinearity, specifically, when the variance of each variable is near equal, $\mathbf{\Sigma}$ yields a small determinant. Below is an example of the latter case.

$$\det \mathbf{C} = \begin{vmatrix} 4 & 3.99 \\ 3.99 & 4 \end{vmatrix} = 4 * 4 - 3.99 * 3.99 = 0.08$$

In summary, we obtain a non-positive definite matrix when there is one variable with constant variance or two variables with perfect multicollinearity. Additionally, the determinant will approach zero as variance decreases for at least one variable or as collinearity increases in certain cases between two variables. Because of these properties, the determinant of $\mathbf{\Sigma}$ can be useful for identifying low variance or high multicollinearity.

Wilks (1932) first proposed using the determinant of $\mathbf{\Sigma}$ as a measure of overall variance for multivariate data terming the measure generalized variance (GV). Wilks described the determinant as a measure of multidimensional scatter, and further applications of the measure include use as a measure of estimator efficiency (Isaacson 1951; SenGupta & Vermeire 1986). Generalized variance is a concept that can be understood visually as the magnitude of the scatter of points in a multidimensional space, with higher GV indicating greater magnitude of scatter and lower GV indicating lesser magnitude of scatter. Using covariance matrix **D**, we obtain the multidimensional scatter plot in **Figure 1.1** with a GV of 3.68:

Figure 1.1: Three dimensional scatter plot where the generalized variance is 3.68. Variance is moderate and points are scattered well across all axes.

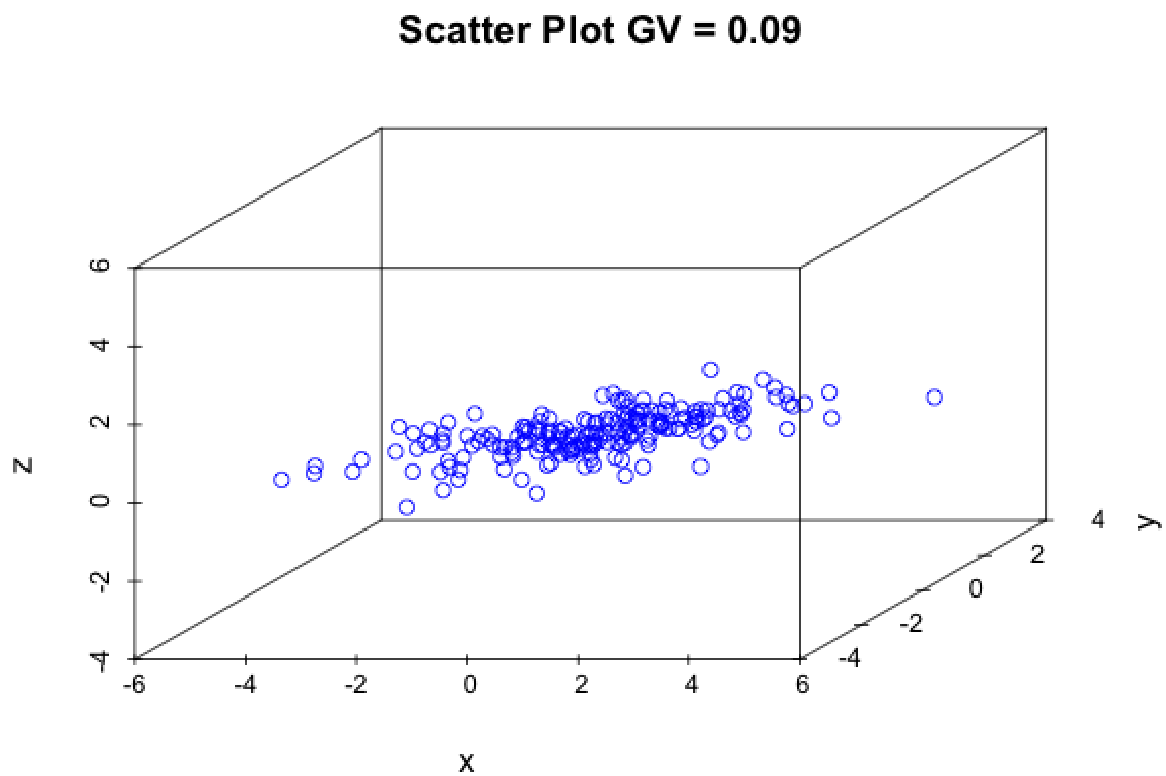


$$\mathbf{D} = \begin{bmatrix} 2.44 & 0.74 & 0.43 \\ 0.74 & 1.45 & 0.26 \\ 0.43 & 0.26 & 1.32 \end{bmatrix}$$

With covariance matrix $\mathbf{E}$ where $\mathbf{E}_{33}$ has low variance, we obtain the multidimensional scatter plot in **Figure 1.2**.

$$\mathbf{E} = \begin{bmatrix} 2.44 & 0.74 & -0.01 \\ 0.74 & 1.45 & -0.01 \\ -0.01 & -0.01 & 0.03 \end{bmatrix}$$

Figure 1.2: Three dimensional scatter plot where the generalized variance is 0.09. Points are clustered more tightly along z axis.



In **Figure 1.2**, the points are much more tightly clustered than in **Figure 1.1**. In sum, the GV, or the determinant of Σ, is a useful measure for understanding how much overall variance exists in multidimensional matrices, with a GV closer to zero being indicative of very little multidimensional variance.

1.2.2 Haitovsky's statistic for high multicollinearity

There is precedent for using the determinant of a matrix as an overall measure of some feature for multivariate data. The issue of very high multicollinearity presents similar problems in the correlation matrix as very low variance (and certain cases of multicollinearity) does in the covariance matrix. A correlation matrix where variables are highly multicollinear with each other yields a determinant very close to zero. Similar to how the inverse of the covariance matrix yields the precision matrix, the inverse of a correlation matrix yields a matrix of regression coefficients where the diagonal is the variance inflation factor (VIF). A correlation matrix with high multicollinearity results in inflated standard errors for the regression coefficients. Rockwell (1975) argues that not only do the inflated standard errors lead to doubt of the results, but we also cannot judge the multicollinearity of a correlation matrix by just looking at the correlation matrix itself. Rockwell argues for the use of the Haitovsky statistic (1969), a measure of high multicollinearity, demonstrating the importance of such a diagnostic by showing that in published papers, variables with high multicollinearity were either mistakenly removed or not removed from analyses despite Rockwell's reassessment using the Haitovsky statistic.

The Haitovsky statistic transforms the determinant of the correlation matrix, yielding an adjusted Chi-square statistic to determine whether the determinant is different from zero. The transformation of the determinant of the correlation matrix is defined below.

$$\chi_h^2(v) = k(log_e)(1 - |\rho|) \tag{1.3}$$

where

$$k = [1 + (2p + 5)/6 - N]$$

where v is the degrees of freedom, $k*(log_e)$ is an adjustment factor defined for Chi-square statistics (Bartlett 1954), $|\rho|$ is the determinant of the correlation matrix, p is the number of explanatory variables, and N is the sample size. The resulting Haitovsky statistic approximately follows a Chi-square distribution. The null hypothesis for the test is that the determinant is zero. We note here that testing a null hypothesis of zero is unrealistic but useful for understanding if the Haitovsky statistic is very *close* to zero. This is in line with the idea of the perinull hypothesis, where, in the case of the Haitovsky test, the value of zero is the limit that is ”least favorable” (i.e., noninvertible) and values approaching zero are unfavorable (i.e., result in high standard errors) (Williams, Jones, & Tukey 1999).

Utilizing the concept in the Haitovsky diagnostic of comparing the determinant of a matrix to zero, we can do the same with the covariance matrix by standardization of the GV. A major difference with using the GV as a diagnostic is that the determinant of the covariance matrix is based on all elements in a matrix, while the Haitovsky test focuses on off-diagonal elements since its main concern is collinearity between variables. Using the GV as a diagnostic, we are concerned with levels of variance and implicitly covariance.

1.2.3 Standardizing generalized variance

With the Haitovsky diagnostic for high multicollinearity, a log-transformed determinant is used to obtain a Chi-square statistic for comparison with zero in the Chi-square distribution. The transformation is needed because the determinant of the correlation matrix itself is not Chi-square distributed, and putting it in a standard distribution makes it possible to test against zero. Additionally, because the determinant is from a correlation matrix, the determinant is bounded, meaning it can only be a value between 0 and 1, with 1 representing completely orthogonal data in the case of the correlation matrix. Because the determinant is already in this standardized format, and the determinant is transformed to belong in a chi-square distribution, it can be tested with a perinull hypothesis of zero.

Similarly, the determinant of the covariance matrix needs to be in a standardized format, preferably one that is distributed in a commonly used distribution like the Chi-square distribution. There are proofs showing how the log GV is normally distributed; however, even then, our determinant would still be unbounded. Simply having a larger number of variables or a variable with a large scale in the dataset would yield a higher GV, regardless of there being variables with very low variance in the data set. Work by SenGupta (1987b), however, allows for a transformation of the determinant that is standardized to the number of variables and Chi-square distributed, making the statistic both bounded and drawn from a familiar distribution in the null case.

SenGupta presents the SGV (1987b; 1987a), a standardization of the determinant, as a way for comparing the overall variance in multivariate data between two populations. The SGV has been used in many applications similar to the GV; however, the SGV allows for comparison of the overall variance of populations with different dimensions. In particular, the SGV has been used for assessing the overall variability of time series of different dimensions, a data type where missing data and thus, loss of a dimension is common (Behara 1983). The formula for SGV is:

$$SGV = |\mathbf{\Sigma}|^{1/p} \tag{1.4}$$

where *p* is the number of variables and $\mathbf{\Sigma}$ is the covariance matrix. In other words, we simply take the *pth* root of the determinant of the covariance matrix. SenGupta showed that the ratio of SGV for one population to SGV for another population is F-distributed (2011), thus, the distribution of the SGV for one population is Chi-square distributed (Kutner et al. 2005). The degrees of freedom are $p(n-p)$, again where p is the number of variables and n is the sample size (SenGupta 1987b). Looking at the degrees of freedom, we encounter a problem with testing the perinull hypothesis that the SGV is zero. We see that even data sets with small sample sizes and only a few variables will have a large degrees of freedom. Just two variables and a sample size of 30 results in 56 degrees of freedom. This makes testing the perinull hypothesis that the SGV is zero intractable since datasets would need to have a very high SGV in order to reject the perinull hypothesis.

An alternative to hypothesis testing is bootstrapping (SenGupta & Ng 2011). With this method, we do not have to be concerned with the shape of the distribution. Instead, we take multiple samples with replacement from our current sample to determine the shape of the distribution of the SGV. For the purposes of diagnosing low variability, we would ask if a 95 percent confidence interval on the bootstrapped sample includes zero. If the confidence interval includes zero, we would conclude that the variability of one of our variables in the data set is too low. Because we are essentially asking if we draw a constant variable in enough bootstrapped samples so that the confidence interval includes zero, we need not distinguish between SGV and GV at this point. Both would lead to the same inference, so from here on, we will refer solely to the GV.

A potential issue with bootstrapping for the purpose of detecting low variability is that we essentially have a combinatorics problem where the question is: In how many bootstrapped samples with replacement do we draw a constant variable? Based on probability, there may be a set limit of differing observations that result in greater than 2.5 percent of resamples containing a constant variable, regardless of sample size. In that case, the test would be unnecessary and we would simply have a cutoff point to be used across all sample sizes.

1.2.4 Distinguishing low variance from multicollinearity

As previously discussed, there are some cases of normal variability where we may obtain a confidence interval where the GV is zero due to high multicollinearity. To address this, we can use both the covariance and correlation matrices to distinguish low variance from multicollinearity. When we have a data set where one or more variables has low variance, we will detect low variability with the GV. In situations where we have low variance, the variables with low variance will usually have very small correlations with other variables, thus, in those cases, high multicollinearity is not a problem. The determinant of such a correlation matrix will be approaching 1, while the determinant of the covariance matrix, or GV (as well as the SGV), will be approaching 0.

In some instances, we will both obtain a confidence interval of the GV that contains zero and fail to reject the perinull hypothesis for the Haitovsky diagnostic. This occurs when the parts of the matrix associated with covariance and the parts of the matrix associated with variance are close to equality, which we explained previously in matrix **C**.

Essentially, when the covariance and the variance are close to equality, most of the variance

Table 1.1: Hypothesized outcomes of GV and Haitovsky statistic

Variance	Collinearity	GV	Haitovsky
Low	High	Detect low variability	Fail to reject
Normal	Normal	Do not detect low variability	Reject
Low	Normal	Detect low variability	Reject
Normal	High	Do not detect low variability (unless equal σ^2 holds)	Fail to reject

is accounted for by relationships between variables. Except for those special cases, we can distinguish low variance from high multicollinearity with both the GV and the Haitovsky diagnostic. For this reason, we cannot only test for low variability on the covariance matrix but must also test for very high multicollinearity on the correlation matrix. *Table 1.1* summarizes expectations for the results of bootstrapping the GV and testing the perinull hypothesis of the Haitovsky statistic, which we test in the simulation.

1.2.5 Gini coefficient

In addition to demonstrating the GV as a measure of variability, we will also investigate the Gini coefficient (Gini 1921) as a measure of variability. The Gini coefficient has long been used as a measure of inequality (Gini 1921; Biewen 2002; Kerm n.d.), and it has been recommended that the Gini coefficient be taught as a measure of variability instead of variance or standard deviation when dealing with categorical data (Kader & Perry 2007), making it potentially ideal for detecting low variability. The Gini coefficient asks: what is the probability of drawing two observations with the same response? The formula is:

$$G = \frac{1}{2\bar{X}N(N-1)} \sum_{i=1}^{N} (2i - N - 1)X_i \tag{1.5}$$

where *N* is the number of observations and X is the number of levels for the given variable. When it comes to making inferences from the Gini coefficient, bias-corrected, bootstrapped confidence intervals are recommended (Dixon et al. 1987) and are what we use here to detect low variability (confidence interval including zero) with the Gini coefficient. There is potential that we will experience similar issues using bootstrapping as we do with the GV, where the results do not depend on sample size. Again, in that case, a cutoff point would be more useful than a sample-size dependent test.

1.3 Aims

This project aims to address the following questions:

1. Can GV, along with the Haitovsky test, help us to differentiate between low variability and high multicollinearity?

2. Can GV and the Gini coefficient be used to detect low variability with high specificity?

3. How do GV and the Gini coefficient perform as the number of variables in the matrix increases?

4. What happens to the standard errors of the regression coefficient for a variable when that variable has extremely low variability as defined by GV and the Gini coefficient?

5. What is the impact of a low variability variable on the standard errors of the regression coefficient of a normal variability variable?

To address Aims 1, 2, 4, and 5, we first conduct a simulation where we model linear regression with two predictors. The two predictors were simulated under varying levels of sample size, variability, strength of relationship with the outcome variable, and correlation between the predictors. To address Aim 3, we use empirical data, varying the number of variables included in the analysis.

2 METHOD

2.1 Simulation Study

In order to assess whether GV and the Gini coefficient can detect a variable with low variability, whether the GV can distinguish low variability from high multicollinearity, and how standard errors perform with the presence of a low-variability variable, we conducted a computer simulation. Even though the impact of removing a variable on the standard errors of another variable can be determined analytically (Kutner et al. 2005), demonstrating the impact across a variety of situations allows for a better understanding of when standard errors will be inflated. Data generation, analysis of simulated data, and analysis of outcomes were all conducted in R. For each condition, 1000 data sets, each with three variables (two predictors and one outcome) of either n=100 or n=1000 were generated. Conditions varied depending on amount of variability in each variable, level of collinearity between the two variables, and strength of relationship between the predictor variables and outcome variable.

Using the ”GenOrd” package in R, the two predictor variables (X_1 and X_2) were generated on a 6-point Likert-type scale. X_1 and X_2 were defined as having either low, medium, or high variability. Marginal cumulative distributions for each level of variability are defined in *Table 2.1*. For the low variability condition, we wanted the number of observations differing from the mode of the sample to center around 4. For this reason, the n=100 and n=1000 conditions had differently

Table 2.1: Cumulative marginal distributions for low, medium, and high variability

Variability	$P(X_i = 1)$	$P(X_i = 2)$	$P(X_i = 3)$	$P(X_i = 4)$	$P(X_i = 5)$	$P(X_i = 6)$
Low (n=100)	0.960	0.990	0.9925	0.9950	0.9975	1
Low (n=1000)	0.996	0.999	0.99925	0.9995	0.99975	1
Medium	0.500	0.600	0.700	0.800	0.900	1
High	0.167	0.333	0.500	0.667	0.833	1

defined cumulative marginal distributions. There were 5 conditions for levels of variability: (1) two variables both with low variability, (2) two variables both with medium variability, (3) one variable with low and one variable with medium variability, (4) one variable with low and one variable with high variability, (5) one variable with medium and one variable with high variability (See *Table 2.1* for definitions of levels of variability). These conditions were chosen in order to determine if low variability variables can still be detected in the presence of medium and high variability variables. Additionally, the conditions where there are no low-variability variables will demonstrate that we will not obtain false negatives using this test.

The correlation between X_1 and X_2 was either low or high. Low correlation was defined as being concentrated around $r = 0.07$. This value was chosen as it is the highest possible threshold for the correlation of categorical data when we have $n = 1000$ and two low-variability (as defined by our simulation) variables (Barbiero & Ferrari 2017). High correlation was defined as being concentrated around 0.9. Because the GenOrd method of simulation does not always reach the desired level of correlation, for the low correlation condition, we removed samples simulated with an absolute value of *r* above 0.2, and for the high correlation condition, we removed samples simulated with an absolute value of *r* below 0.8. These conditions were chosen to ensure that we can distinguish between low variability and high multicollinearity as well as to ensure that we can determine when variables both have low variability and are highly correlated with each other.

Finally, an outcome variable, *y*, was generated. A regression model was specified where the regression coefficient for X_1 was 0.4, and the regression coefficient for X_2 (the lower variability variable) was either 0.4 or 0.1. The standard deviation of *y* was 0.4.

Taken together, there were a total of thirty-two conditions which are summarized in *Table 2.2*. The conditions were not fully crossed, as it is not possible to simulate a high correlation in some of the conditions (e.g. X_1 with medium and X_2 with low variability) (Barbiero & Ferrari 2017).

2.1.1 Hypotheses

In this simulation, we had several outcomes of interest:

Table 2.2: Simulation conditions

Condition	Variability X_1	Variability X_2	Correlation	β_{X_1}	β_{X_2}	n
1	Low	Low	Low	0.4	0.4	100
2	Low	Low	Low	0.4	0.1	100
3	Low	Low	High	0.4	0.4	100
4	Low	Low	High	0.4	0.1	100
5	Medium	Low	Low	0.4	0.4	100
6	Medium	Low	Low	0.4	0.1	100
7	High	Low	Low	0.4	0.4	100
8	High	Low	Low	0.4	0.1	100
9	Medium	Medium	Low	0.4	0.4	100
10	Medium	Medium	Low	0.4	0.1	100
11	Medium	Medium	High	0.4	0.4	100
12	Medium	Medium	High	0.4	0.1	100
13	High	Medium	Low	0.4	0.4	100
14	High	Medium	Low	0.4	0.1	100
15	High	Medium	High	0.4	0.4	100
16	High	Medium	High	0.4	0.1	100
17	Low	Low	Low	0.4	0.4	1000
18	Low	Low	Low	0.4	0.1	1000
19	Low	Low	High	0.4	0.4	1000
20	Low	Low	High	0.4	0.1	1000
21	Medium	Low	Low	0.4	0.4	1000
22	Medium	Low	Low	0.4	0.1	1000
23	High	Low	Low	0.4	0.4	1000
24	High	Low	Low	0.4	0.1	1000
25	Medium	Medium	Low	0.4	0.4	1000
26	Medium	Medium	Low	0.4	0.1	1000
27	Medium	Medium	High	0.4	0.4	1000
28	Medium	Medium	High	0.4	0.1	1000
29	High	Medium	Low	0.4	0.4	1000
30	High	Medium	Low	0.4	0.1	1000
31	High	Medium	High	0.4	0.4	1000
32	High	Medium	High	0.4	0.1	1000

1. **Proportion of samples where we detect low variability using the GV and high multicollinearity using the Haitovsky statistic.** We expected detection of low variability and high multicollinearity to align with the following hypotheses. We hypothesize that by bootstrapping the GV, we will detect when one or more of the variables has low variability. That is, our confidence intervals will include zero. In conditions where we have two normal variability variables, we expect that our confidence intervals will also include zero because the variables will have equal variance as well as a high correlation. In the remaining conditions, we hypothesize that we will not detect low variability. Regarding multicollinearity, we hypothesize that the the Haitovsky test will detect multicollinearity in the conditions with high correlation coefficients. That is, we will fail to reject the perinull hypothesis that the determinant of the correlation matrix is zero when we generate data to have a high correlation, and we will reject the perinull hypothesis that the determinant of the correlation matrix is zero when we generate data to have a low correlation.

2. **Proportion of samples where we detect low variability using the Gini coefficient.** We expect that the Gini will follow the same pattern as the GV, where with low variability conditions, we will detect low variability. However, because the Gini is calculated for each individual variable, meaning that multicollinearity will not factor into the results, we do not expect conditions with a high correlation coefficient but medium variability to result in our confidence intervals including zero.

3. **Standard errors of X_2.** We hypothesize that the standard errors of β for X_2 will be larger when the condition for X_2 is low variability and when the correlation between X_1 and X_2 is high ($r > 0.8$).

4. **Standard errors of X_1.** We hypothesize that the standard errors of β for X_1 will be larger when the condition for X_2 is low variability and when the correlation between X_1 and X_2 is high.

2.2 Empirical Example

To demonstrate how the GV performs with varying numbers of variables in the model, we provide an empirical example using daily diary data from a study of individuals diagnosed with personality disorders. In this study, individuals reported daily emotional and behavioral data. Of the behavioral variables, we focus on the problem of low variability in reporting drug use. If someone infrequently uses drugs, they may only report drug use on a few days out of a 100-day daily diary study. It may also be that frequent drug users only report not using drugs on a few occasions. If this variable is of theoretical interest to a researcher, it may be of importance to determine if the drug-use variable can be used in their model.

We use these data to demonstrate a real-world example of how bootstrapping the GV will perform. There are four conditions in the empirical example. We began by obtaining the bootstrapped confidence interval of the GV on a 2 variable example and follow with 5, 15, and 30 variable examples to determine if the GV holds up at increasing numbers of variables (p). Variables are selected based on level of variability and normality. Each condition has one variable with very low variability. The remaining variables have moderate to high variability and are the most normally distributed variables in order to avoid confounding issues such as skewness. Additionally, variables are not included if there is a correlation between two variables greater than 0.6. This is to avoid confounding issues of multicollinearity.

The outcome of interest is the proportion of times we detect low variability using the GV. Because the Gini coefficient is calculated independent of other variables in the model, we do not test how it performs under increasing numbers of variables. We expect the diagnostic to perform well when p=5 and p=15 but expect some detriment at p=30.

3 RESULTS

In all, 32,000 sets of data were simulated and modeled. Sets of data where the correlation between X_1 and X_2 was between 0.2 and 0.8 were removed, as well as cases where there was perfect multicollinearity between X_1 and X_2 or one of the variables was generated with variance of zero. This left a total of 28,845 sets of data. The distributions of correlation coefficients for the remaining data sets are displayed in **Figure 3.1a-b**. The low correlation condition was bimodal with a peak centering around $r = 0$ and $r = 0.07$. The distribution for the high correlation condition was leptokurtic centering around $r = 0.9$.

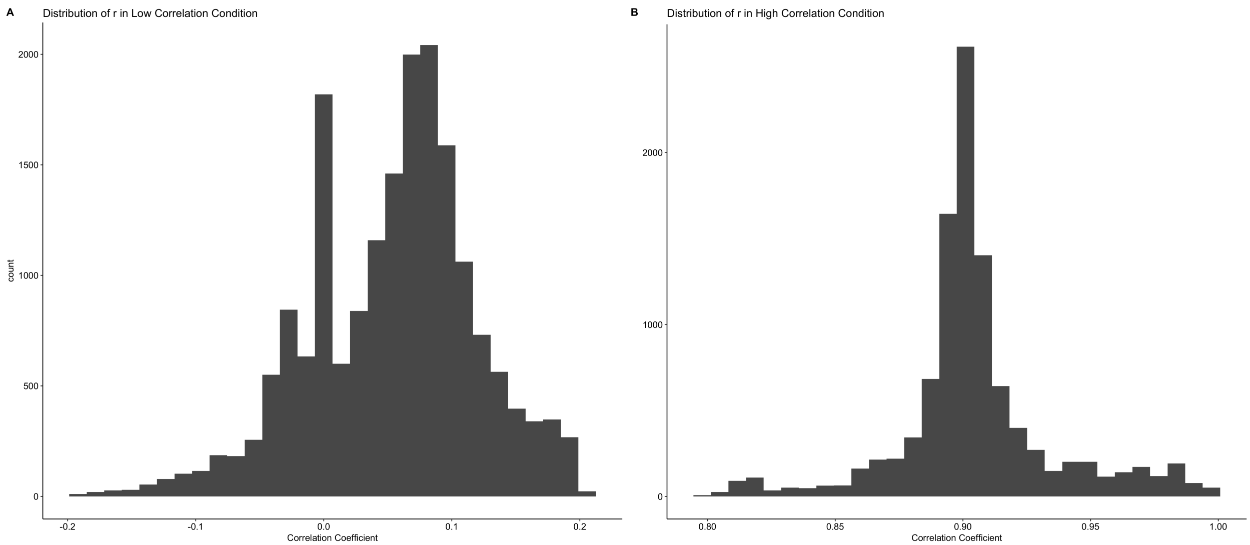
3.1 Performance of the GV, the Gini coefficient, and the Haitovsky test

The overall pattern of detecting low variability and high multicollinearity for the indicators examined was as hypothesized. In conditions when one or more variables had low variability, we often obtained confidence intervals of the GV and the Gini coefficient that included zero, when multicollinearity was high, we retained the perinull hypothesis that the Haitovsky statistic is zero, and in cases with high multicollinearity but medium to high variance, we obtained confidence intervals of the GV that included zero. A heatmap displaying proportions of detection by each diagnostic can be found in **Figure 3.2**. While the low correlation condition generated datasets with up to twelve observations differing from the mode, here we only look at cases with up to seven observations differing from the mode since there is no detection of low variability by the Gini coefficient or the GV when 6 or more observations are differing from the mode (unless we consider zero-inclusive confidence intervals of the GV in cases of high multicollinearity).

3.1.1 Thresholds of low variability detection

Looking at the heat map, we see that regardless of the level of correlation between X_1 and X_2 or the sample size, both the GV and the Gini coefficient detect low variability when only one or two observations differ from the mode. The GV detects low variability, regardless of correlation

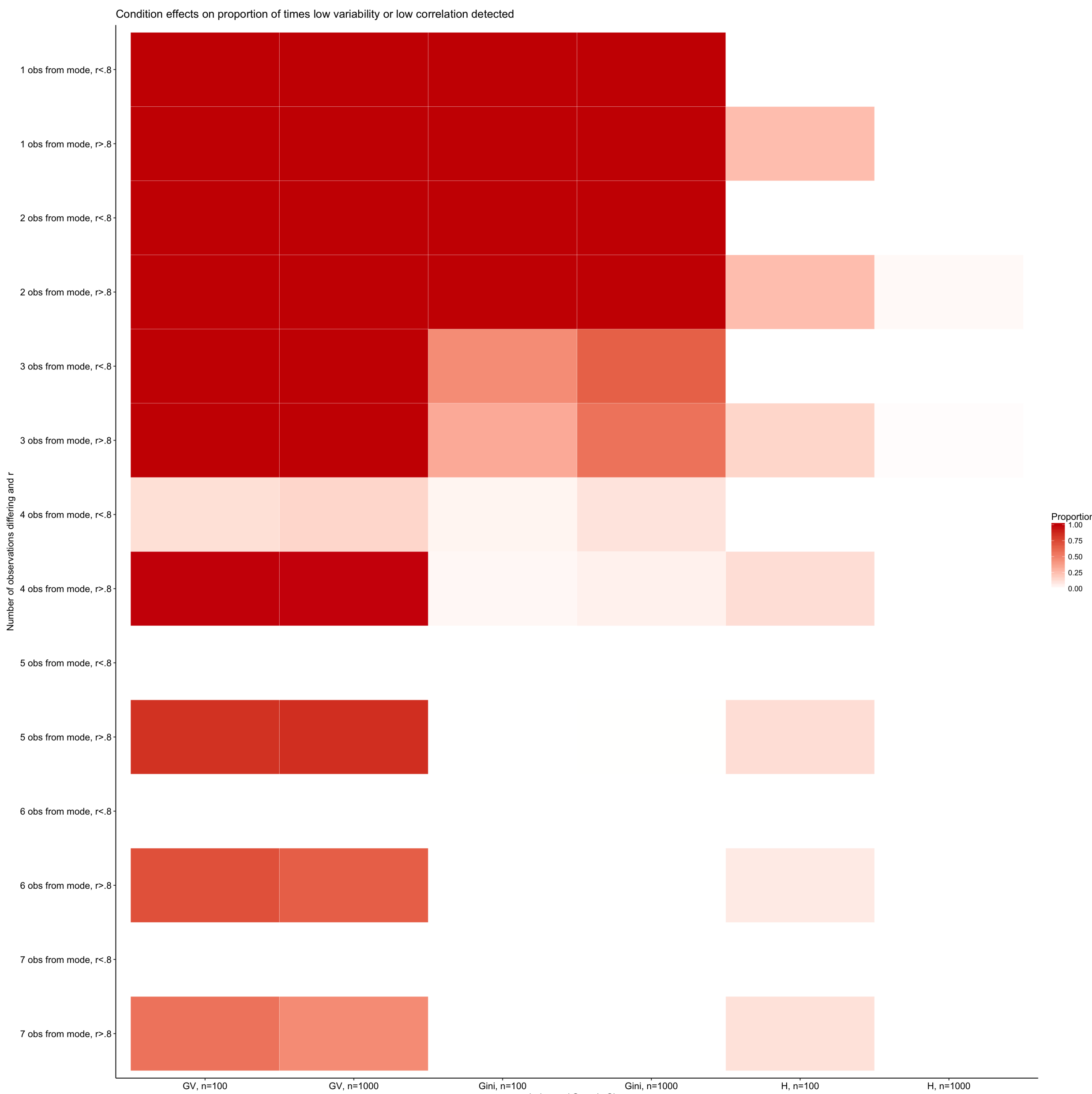
Figure 3.1: Distribution of correlation coefficients generated by simulation in low and high correlation conditions.



or sample size in nearly all cases where three observations differ from the mode (When the sample size is 1000, and the correlation is low, 99.9 percent of cases are detected). Once we reach four observations differing from the mode, whether we detect low variability with the GV depends on the level of correlation. In only a few cases (12 and 16 percent for n=100 and n=1000, respectively) do we detect low variability when the correlation is low and four observations differ from the mode. With five or greater observations differing from the mode, we only detect low variability when the correlation is high. These findings align with our expectation that the GV would also pick up cases of high multicollinearity and that bootstrapping the GV might be sample size independent.

The test of the Gini coefficient evaluates each variable independent of the other, so level of correlation did not have any impact on detection of low variability as we observed with the GV. When three or more observations differ from the mode, we observe fewer cases where we detect low variability. Additionally, there are slightly more cases where we detect low variability when n=1000 versus n=100. When five or more observations differ from the mode and n=100, we always detect low variability. When n=1000, we detect low variability in 0.09 percent of cases. When six or more observations differ from the mode, we never detect low variability, regardless of sample size.

Figure 3.2: Heat map displaying the proportion of cases where we detect low variability or low correlation by number of observations varying from the mode for X_1 or X_2, the correlation between X_1 and X_2, and sample size.



3.1.2 Multicollinearity

As expected, the Haitovsky test did not detect multicollinearity in any of the samples from the low correlation condition. Detection of multicollinearity varied by sample size. When $n = 1000$, we only detected high multicollinearity in a few cases when there were two or three observations differing from the mode. When $n = 100$, there were more cases where we detected high multicollinearity, and detection of high multicollinearity occurred in all variability conditions. It makes sense that we detect high multicollinearity with larger sample sizes since we would have greater power to reject the perinull hypothesis. Finally, based on the simulation, the Haitovsky test detects fewer cases of high multicollinearity than bootstrapping the GV. As noted above, we frequently obtained confidence intervals of the GV that included zero in conditions where $r > .8$, regardless of the amount of variability in the sample. Detection of high multicollinearity occurred at a higher rate for the GV than for the Haitovsky test, suggesting that the Haitovsky test is more discriminating when it comes to diagnosing high multicollinearity.

3.2 Standard errors of X_1.

To assess the impact of low variability on the standard errors of other variables in the model, we looked at how the standard errors of X_1 change when X_2 is excluded from the model in cases where X_2 was identified as having low variability by the GV and the Gini coefficient. Cases where we failed to reject the perinull hypothesis (that the determinant is zero) for the Haitovsky test were excluded to prevent conflation with multicollinearity. Box plots in **Figure 3.3** and **Figure 3.4** show the distribution of the standard errors of the regression coefficient of X_1 when X_2 (low variability variable) is included in the model and when X_2 is excluded from the model and when the regression coefficient of X_2 is either significant or non-significant. Panel A for each figure shows the pattern of standard errors for cases where low variability was identified using the GV, and Panel B for each figure shows the pattern of standard errors for cases of low variability identified using the Gini coefficient. In **Figure 3.3**, we display the impact on standard errors when X_1 is generated to have medium or high variability, and in **Figure 3.4**, we display the impact on standard errors when X_1 is generated to have low variability.

Figure 3.3: Box plots showing distribution of standard errors of X_1 when X_1 is generated to have normal variability and when X_2 is identified as having low variability.

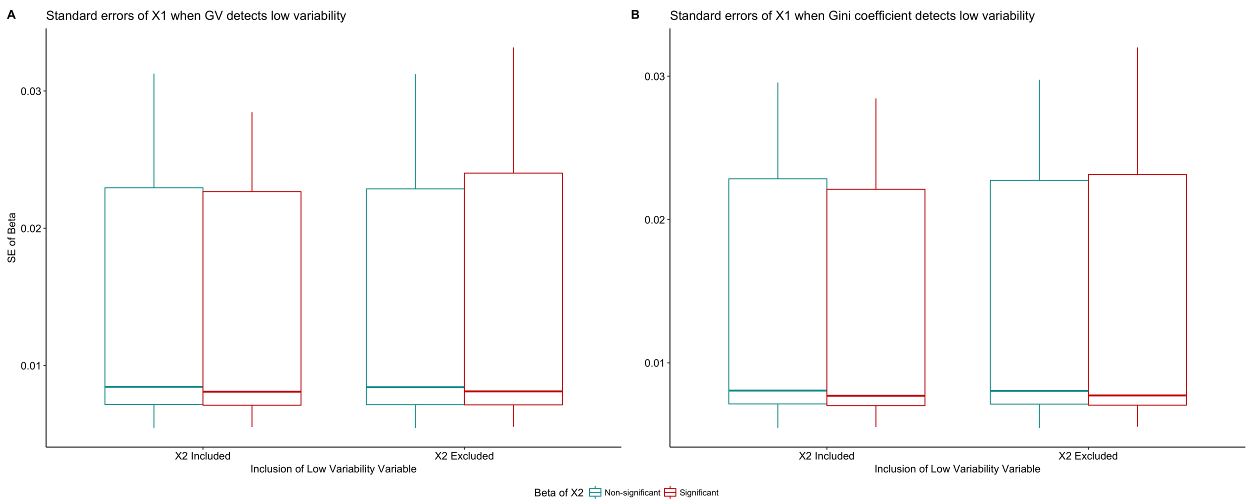
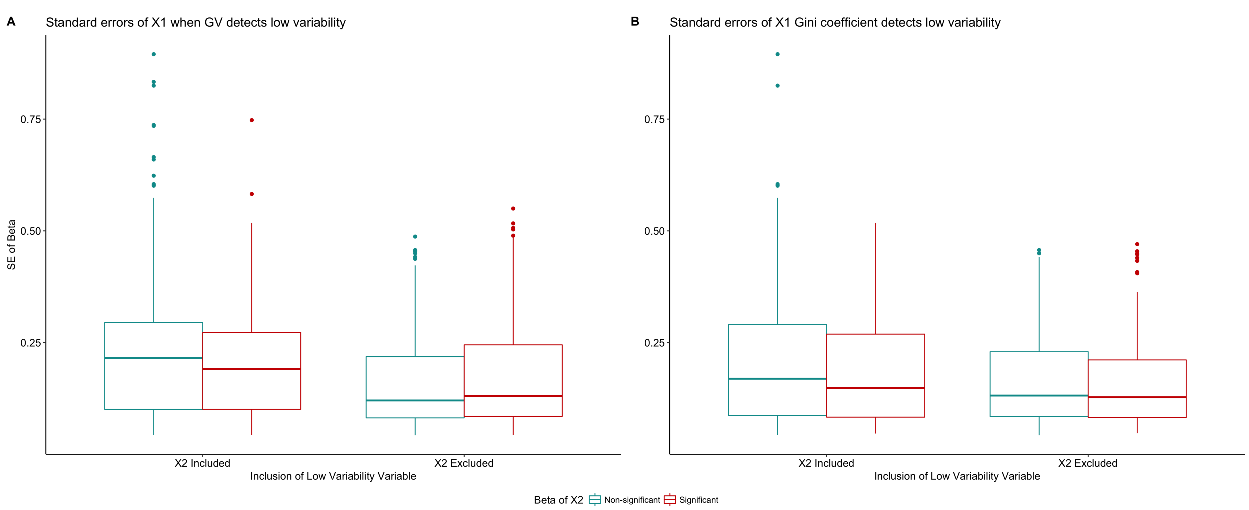


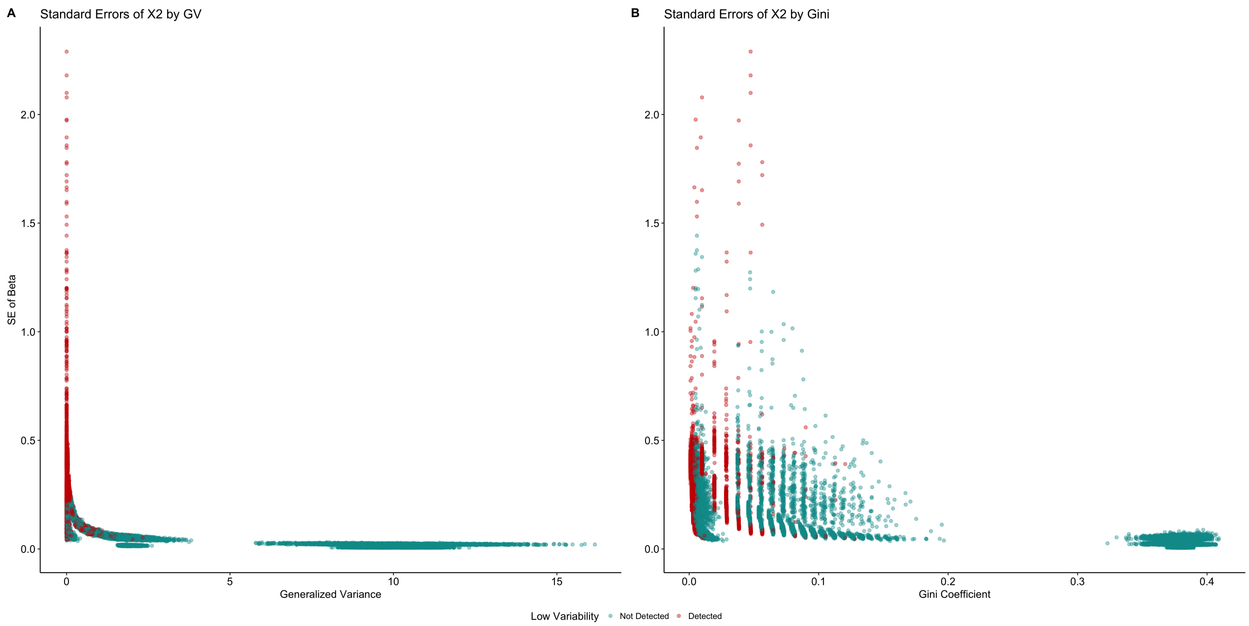
Figure 3.4: Box plots showing distribution of standard errors of X_1 when X_1 is generated to have low variability and when X_2 is identified as having low variability.



In **Figure 3.3a** and **Figure 3.3b**, we see that the GV and the Gini coefficient have similar patterns of standard errors depending on whether we exclude or include a low variability variable in the model and depending on whether the low variability variable has a significant or non-significant regression coefficient. Two, two-way factorial ANOVAs were conducted corresponding with the boxplots in **Figure 3.3**. When looking at cases where the GV detected low variability (Figure 3.3a), there were no significant main or interaction effects ($F(3, 6146) = 1.42$, $p = 0.24$). However, when looking at cases where a test of the Gini coefficient detected low variability (Figure 3.3b), there were significant differences ($F(3, 5180) = 5.17$, $p < 0.01$). Tukey's HSD test revealed a main effect where the standard errors of X_1 were significantly smaller when X_2 was significant ($p < 0.01$, $\eta^2 = 0.003$).

Next, these tests were repeated for cases where X_1 was generated to have low variability and X_2 was identified as having low variability by the GV and the Gini coefficient. From the box plots in **Figure 3.4**, we see that there is greater variability in the standard errors of X_1 when X_1 is generated to have low variability versus medium or high variability (as displayed in **Figure 3.3**). Looking at cases where the GV detected low variability in X_2 (Figure 3.4a), a two-way factorial ANOVA revealed significant differences ($F(3, 3320) = 64.94$, $p < 0.01$). Tukey's HSD test revealed a main effect based on whether X_2 was included in the model, where excluding X_2 from the model resulted in smaller standard errors of X_1 ($p < .01$, $\eta^2 = 0.04$). There was also a significant interaction effect ($p < .01$, $\eta^2 = 0.006$), where the standard errors of X_1 are lower when X_2 is included in the model and significant but higher when X_2 is included in the model and non-significant. When X_2 is excluded from the model, the standard errors of X_1 are higher if X_2 was significant but lower if X_2 was non-significant. Now looking at cases where the Gini coefficient detected low variability (Figure 3.4b), a two-way factorial ANOVA revealed significant differences ($F(3, 1252) = 11.17$, $p < 0.01$). Tukey's HSD test revealed a main effect based on whether X_2 was included in the model, where excluding X_2 from the model resulted in smaller standard errors of X_1 ($p < .01$, $\eta^2 = 0.02$). Unlike with cases where the GV identified low variability in X_2, there was no interaction effect when the Gini coefficient detected low variability in X_2.

Figure 3.5: Scatter plots of the standard errors of X_2 correlated with the values of the GV and the Gini coefficient Points in red indicate that the given index identified the variable as having low variability, while points in green indicate that the variable did not have low variability.



3.3 Standard errors of X_2.

We also looked at the standard errors of X_2 (with X_1 included in the model) to see if the GV and the Gini detect when standard errors of the regression coefficient for a variable will be larger due to low variability. **Figure 3.5** shows scatter plots of the GV and the standard errors of X_2 and the Gini coefficient and the standard errors of X_2. We see that with both indices, a lower value of the GV or Gini coefficient is associated with higher standard errors. Based on the scatter plots, the GV appears better at detecting when the standard errors will be higher. With the Gini coefficient, there are cases where the standard errors appear high, but we still do not detect low variability. This outcome may be due to the fact that the Gini coefficient is not taking the other variable in the model into account. Based on our above results, the standard errors of X_2 may be higher when X_1 has low variability or when there is high multicollinearity, which the Gini coefficient will not detect.

We conducted two t-tests to determine if there were significant differences in the standard errors depending on if each diagnostic detected low variability. We found that the standard errors were significantly higher ($t(8381) = -101.32$, $p < 0.01$, $d = -1.34$) when the GV detected low variability ($mean = 0.24$) versus when the GV did not detect low variability ($mean = 0.05$). We found similar results when comparing standard errors of cases where the Gini coefficient detected low variability ($mean = 0.27$) versus when the Gini coefficient did not detect low variability ($mean = 0.07$) ($t(4821) = -75.87$, $p < 0.01$, $d = -1.86$).

3.4 Empirical Example

Of the 112 participants in the dataset, we selected one participant who reported drug use on all 101 days of reporting except three. This allows us to test the performance of the GV at the threshold of detection we found in our simulation study, where all cases were found to have low variability according to the test of the GV when three observations differed from the mode. When four observation differ from the mode, only some cases are detected, so it would be difficult to determine if any changes in detection are due to differing the number of variables in the model.

All correlations between variables included were below 0.6. All variables, except for the drug use variable, had moderate variability. Low variability was detected in every condition (p=2, p=5, p=15, and p=30). This indicates that bootstrapping the GV is effective in detecting low variability when there are varying numbers of variables in a model, offering support for our claim that we do not need to take into account the number of variables when using the bootstrap method.

4 DISCUSSION

4.1 Thresholds of exclusion

To summarize, we found that regardless of sample size, we will always detect low variability with the GV (when correlation is low) or the Gini coefficient when at least six or more observations in a sample differ from the mode. This aligns with our expectation that, using bootstrapping, where we resample with replacement, the probability of selecting a large enough proportion of samples (2.5 percent in the case where $\alpha = 0.05$) where only the mode was sampled reaches zero at a certain threshold (six or more observations differing from the mode). That this occurs regardless of sample size indicates that suggesting a cutoff point of five could be more useful than suggesting a diagnostic where whether the bootstrapped statistic includes zero comes into question when three, four, or five observations differ from the mode (in the case of the Gini coefficient) or four observations differ from the mode (in the case of the GV). Essentially, we would not need to test for low variability if at least five observations differed from the mode. However, a cutoff point would discourage an understanding of when low variability is actually a problem for estimation of a model.

4.2 Impact on model

Our results concerning the inclusion and exclusion of a low-variability variable (as defined by the GV or the Gini coefficient) in a model with a normal-variability variable indicate that there is not an impact of including a low-variability variable on the standard errors of a normal-variability variable. We do see problems, however, when we have two low-variability variables in the model. When both variables have low variability, excluding one of the variables from the model may be preferable. Researchers should also consider whether the variable is theoretically meaningful, as excluding a potentially significant variable, is going to result in higher standard errors, even if that variable has very low variability. For this reason, using the Gini coefficient to determine low variability may be preferable. While we observed an interaction effect with the test of the GV where the standard errors were higher when a significant X_2 was excluded from the model, we did not observe this interaction effect when using the Gini coefficient to determine low variability. Standard errors were smaller regardless of the significance of X_2. This indicates that the Gini coefficient, which was more discriminating than the GV, may be better suited for determining if a low-variability variable will result in higher standard errors of beta coefficients in the model.

When looking at the standard errors of the low-variability variable itself (X_2), we also see evidence that the Gini coefficient is more discriminating. While both the GV and the Gini coefficients demonstrated significant differences in the standard errors of X_2 when X_2 was identified as having low variability versus not identified as having low variability, plots of the standard errors by the given diagnostic measure show that we still detect low variability with the Gini coefficient in some cases where the standard error of X_2 is higher than average. It is likely this occurs because the GV is sensitive to cases of higher multicollinearity, while the Gini coefficient only assesses for low variability. Both measures, but more so the GV, do fairly well in detecting when the standard errors of X_2 will be high. This may be useful for researchers who need to take care in interpretation of the results of low-variability variables. When the variability is low, there is still potential to obtain small standard errors; however, because the standard errors can be so unstable at the lowest levels of variability, it may be advisable to interpret one's results with caution.

4.3 Multicollinearity

Based on our results, the Haitovsky test will not allow us to differentiate between low variability and high multicollinearity as we had expected. The Haitovsky test is more discriminating than bootstrapping the GV. There are many cases with five or more observations differing from the mode where we obtain a confidence interval of the GV that includes zero when the correlation is high but not when it is low, suggesting that detection of low variability when five or more observations differ from the mode is due to the level of correlation between X_1 and X_2. If we used the Haitovsky test to differentiate between low variability and high multicollinearity, we would likely determine that the variability is too low, when in fact it is the level of correlation having a larger impact on the GV.

4.4 Future directions

Because the Gini coefficient is more discriminating for determining when standard errors of other low-variability variables will be affected by inclusion of another low-variability variable, is not conflated with multicollinearity, and is somewhat more sensitive to sample size, it may be a better diagnostic tool for judging low variability than the GV. However, the Gini coefficient is still subject to the same problem of bootstrapping as the GV, where we always detect low variability when six or more observations vary from the mode. Determining when the GV or the Gini coefficient will detect low variability analytically (calculating the probability of sampling constant variables when bootstrapping) will be of future use.

Additionally, more exploration into how low variability plays a role in more complicated models is necessary to determine if our finding that the standard errors of normal-variability variables are not affected by inclusion of a low-variability variable in the model remains. Of particular use to intensive longitudinal data with multiple individuals would be determining how low variability in individuals impacts group and individual level estimates.

4.5 Final conclusions

In many cases, low variability will not be a problem for other variables in the model, except when the other variable in the model also has low variability. For this instance, we recommend bootstrapping the Gini coefficient to determine which variables are most likely to impact other variables in the model. It, however, may be just as effective to use five observations differing from the mode as a cutoff point.

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. doi: 10.1002/hbm.22100

Barbiero, A., & Ferrari, P. A. (2017). An R package for the simulation of correlated discrete variables. *Communications in Statistics: Simulation and Computation*, *46*(7), 5123–5140. doi: 10.1080/03610918.2016.1146758

Bartlett, M. (1954). A note on the multiplying factors for various chi square approximations. *Journal of the Royal Statistical Society*, *16*(2), 296–298.

Behara, M. (1983). Generalized variance statistic in the testing of hypothesis in complex multivariate Gaussian distributions. *Archives of Mathematics*, *40*, 538–543.

Biewen, M. (2002). Bootstrap inference for inequality, mobility and poverty measurement. *Journal of Econometrics*, *108*(2), 317–342. doi: 10.1016/S0304-4076(01)00138-5

Chambers, W. (1985). The assessment of affective disorders in children and adolescents by semistructured interview: test-retest reliability of the Schedule for Affective Disorders and. *Archives of General Psychiatry*, *42*(7), 696–702.

Dixon, P. M., Weiner, J., Mitchell-olds, T., Woodley, R., Dixon, P. M., Weiner, J., … Woodley, R. (1987). Bootstrapping the Gini Coefficient of Inequality Linked references are available on JSTOR for this article : NOtes and Com. *Ecology*, *68*(5), 1548–1551.

Fox, J., & Monette, G. (1992). Generalized Collinearity Diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183.

Gibbons, F. D., & Roth, F. P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, *12*(10), 1574–1581. doi: 10.1101/gr.397002

Gini, C. (1921). Measurement of Inequality of Incomes. *The Economic Journal*, *31*(121), 124–126.

Haitovsky, Y. (1969). Multicollinearity in regression analysis: Comment. *The Review of Economics and Statistics*, *51*(4), 486–489.

Isaacson, S. L. (1951). On the Theory of Unbiased Tests of Simple Statistical Hypotheses Specifying the Values of Two or More Parameters. *The Annals of Mathematical Statistics*, *22*(2), 217–234.

Kader, G. D., & Perry, M. (2007). Variability for Categorical Variables. *Journal of Statistics Education*(2), 1–17.

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.

Kerm, P. V. (n.d.). Inference on Inequality Measures : A Monte Carlo Experiment. *Journal of Economics*, 283–306.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). New York, NY: McGraw Hill.

Lebo, M. A., & Nesselroade, J. R. (1978). Intraindividual differences dimensions of mood change during pregnancy identified in five P-technique factor analyses. *Journal of Research in Personality*, *12*(2), 205–224. doi: 10.1016/0092-6566(78)90098-3

Rockwell, R. C. (1975). Assessment of Multicollinearity: The Haitovsky Test of the Determinant. *Sociological Methods and Research*, *3*(4), 308–320. doi: 10.1177/004912417500300304

SenGupta, A. (1987a). Generalizations of Barlett's and Hartley's tests of homogeneity using overall variability. *Communications in Statistics - Theory and Methods*, *16*(4), 987–996.

SenGupta, A. (1987b). Tests for standardized generalized variances of multivariate normal populations of possibly different dimensions. *Journal of Multivariate Analysis*, *23*(2), 209–219. doi: 10.1016/0047-259X(87)90153-9

SenGupta, A., & Ng, H. K. T. (2011). Nonparametric test for the homogeneity of the overall variability. *Journal of Applied Statistics*, *38*(9), 1751–1768. doi: 10.1080/02664763.2010.529876

SenGupta, A., & Vermeire, L. (1986). Locally Optimal Tests for Multiparameter Hypotheses. *Journal of the American Statistical Association*, *81*(395), 819–825.

Stevens, G. V. G. (1998). On the Inverse of the Covariance in Portfolio Analysis. *The Journal of Finance*, *53*(5), 1821–1827.

Stice, E., Shaw, H., & Marti, C. N. (2007, apr). A Meta-Analytic Review of Eating Disorder Prevention Programs: Encouraging Findings. *Annual Review of Clinical Psychology*, *3*, 207–231. doi: 10.1146/annurev.clinpsy.3.022806.091447

Whitley, D. C., Ford, M. G., & Livingstone, D. J. (2000, sep). Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *Journal of Chemical Information and Computer Sciences*, *40*(5), 1160–1168. doi: 10.1021/ci000384c

Wilks, S. S. (1932). Certain Generalizations in the Analysis of Variance. *Biometrika*, *24*(3/4), 471–494. doi: 10.2307/2331979

Williams, V., Jones, L. V., & Tukey, J. W. (1999). Controlling Error in Multiple Comparisons , with Examples from State-to-State Differences in Educational Achievement. *American Educational Research Association*, *24*(1), 42–69.

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.