A Study of Dispersion Effect Identification using Levene-Type Transformations in Replicated Factorial Designs

by

Haley Cornelius

A Thesis
presented to
The University of Guelph

In partial fulfilment of requirements
for the degree of
Master of Science
in
Mathematics and Statistics

Guelph, Ontario, Canada

ABSTRACT

A Study of Dispersion Effect Identification using Levene-Type Transformations in Replicated Factorial Designs

Haley Cornelius
University of Guelph, 2017

Advisors:
Dr. Gary Umphrey

Interest in the analysis of dispersion effects has become increasingly popular and many methods for identifying such effects in factorial designs have been proposed; however, many methods for replicated designs typically result in a loss in degrees of freedom. Levene-type transformations were introduced as a means to transform the response variable such that each observation is now a measure of dispersion. In this thesis, six Levene-type transformations of the response variable in a 2^f factorial design will be performed and analyzed using an analysis of variance to identify dispersion effects. The use of these transformations, compared to a previous method, proves adequate with limitations. It is also of interest to determine how increasing the number of factors while holding the replication size and power level constant affects the detectable effect size. For a fixed power, the decrease in detectable effect size as the number of factors increases is quantified.

To my parents

Acknowledgments

I would first and foremost like to thank my advisor Dr. Gary Umphrey for his knowledge and support throughout the past couple years. Your trust and confidence in my abilities helped to make this thesis possible and is something I will take with me in the future.

I would also like to thank my advisory committee Dr. Jeremy Balka and Dr. Ed Carter for their time and feedback throughout this process, as well as their invaluable teachings in courses I have had the pleasure of taking with them over the years. A big thank you to the entire Statistics department and all the wonderful professors I have been lucky enough to learn from, and to Susan McCormick for your kindness and assistance in the completion of my degree.

Lastly, I would like to thank my friends and family for their love, support and laughter when I needed it most. A special thank you to Emma and Denys for our never ending statistics discussions. Most importantly, thank you to my amazing parents for your unwavering faith in all that I do.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Tests for homogeneity of variance have long been used to verify assumptions of statistical models. They are a common practice in testing assumptions before performing an analysis of location effects, such as an analysis of variance. If the test shows an inequality of variance among groups, typically a transformation of the response variable is conducted in order to satisfy the assumptions and continue with analysis. However, this difference in variances is seldom examined as a primary analysis and is the interest of this thesis. In particular, this thesis focuses on the dispersion effect identification using Levene-type transformations in replicated factorial designs.

Examining dispersion effects became common in the machining industry where determining which factor settings affect the dispersion of an outcome is of interest. For example, it is important to decrease the variation in a production process so that the end product is as consistent as possible. There has also been interest in dispersion in agricultural production, as well as many biological scenarios where variation among populations may exist, such as genetic diversity or species adaptation (Boos and Brownie, 2004).

Similar to the definition of a location effect, a dispersion effect exists when the variation of the response is different among levels of a given factor. These can be

determined for a main dispersion effect of a single factor, as well as for an interaction dispersion effect for a combination of factors.

When there are multiple factors that can be set or observed at a number of levels, a factorial design arises. In this type of design, the location effects are commonly analyzed using an analysis of variance to determine differences in the mean response among the factor levels, but analysis of dispersion effects is less common.

There are many methods currently used to analyze dispersion effects in factorial designs. Some of these methods involve Levene-type transformations of the response variable, which originated from the well-known Levene test for homogeneity of variance. It is these transformations of the response variable, which can be analyzed using an analysis of variance to identify dispersion effects in a replicated factorial design, that are the primary focus of this thesis.

Sample size is also an important aspect of any design and analysis. Although there has been some study of how replication size affects the power of these certain analyses, this has not been examined in conjunction with the number of factors in the design. This change in power as we increase the number of factors will be another focus of this thesis.

The thesis will proceed as follows. Chapter 2 will discuss background and current methods for analyzing dispersion effects. Chapter 3 will propose a simplified method. Chapter 4 will present the methods used for two simulation studies: (1) a comparison to a current method and (2) an analysis of how the number of factors in the design for a given number of replications affects the power of the proposed method. Results of both simulations follow the methods. Chapter 5 provides an

analysis of three different datasets from different fields as applications. Chapter 6 will discuss results and future work.

Chapter 2

Background

2.1 Levene's Test for Homogeneity of Variance

Levene (1960) proposed a test for equality of variance among multiple treatment groups, which has since become a popular test of this kind as it is more robust to non-normality than other well known methods, including the likelihood ratio test, Bartlett's test, and the log-anova test (Conover et al., 1981). It stems from the idea that σ^2 is equal to the expectation of $(X_i - \mu_i)^2$. Levene generalized this idea to

$$W_{ij} = g(|X_{ij} - \bar{X}_i|) \tag{2.1}$$

where $g(X)$ is any monotonically increasing positive function and $\bar{X}_i$ is used in place of μ_i since the population parameter is rarely known. This idea was used to transform each observation and perform a one-way ANOVA among groups using the new response variable, allowing us to maintain the same number of observations. Although the observations are not independent due to the use of $\bar{X}$, Levene suggested that the correlation between $|X_{ij} - \bar{X}_i|$ and $|X_{ik} - \bar{X}_i|$ will have little effect on the distribution of the F statistic (Levene, 1960).

In Levene (1960), a Monte Carlo study to test four different transformations of the response variable is presented and it was found that $z_{ij} = |X_{ij} - \bar{X}_i|$ and $s_{ij} = |X_{ij} - \bar{X}_i|^2$ produced the best power and significance levels, while recommending z_{ij} because of the simplicity of computation.

It is also commonly known that the median can be a better estimate of central tendency for non-symmetric distributions, which led to an extension of Levene's test by Brown and Forsythe (1974). Their method includes the use of the sample median in place of the sample mean, and they also examined the use of a 10% trimmed mean. It was found that the trimmed mean provided better power for symmetric long-tailed distributions, while the median resulted in higher power for skewed distributions (Brown and Forsythe, 1974).

Conover et al. (1981) conducted a thorough comparison of fifty-six different tests for homogeneity of variance in order to determine which tests were robust and most powerful. They defined a test to be robust if the maximum Type I error rate is below 10% for a test at the 5% level of significance. The Brown-Forsythe variation of the Levene test using the median was determined to be one of three superior tests in terms of robustness and power (Conover et al., 1981). Levene's test has continued to be a popular area of research (Schultz, 1985; Gastwirth et al., 2009; Parra-Frutos, 2009).

It was interest in Levene's test for homogeneity of variance that originally motivated this thesis.

2.2 Current Methods for Detecting Dispersion Effects

Interest in the analysis of dispersion effects originally arose in quality control in the manufacturing industry, with the idea that decreasing variability could improve process outcomes (Taguchi, 1986). Since then, various methods for identifying such effects in both unreplicated and replicated designs have been proposed (Nair and Pregibon, 1988; Brenneman and Nair, 2001). The main problem with the analysis for unreplicated experiments is that most methods require that a location model be fit and the residuals of this model can be used for dispersion analysis. This introduces the obvious problem that the analysis of the dispersion effects is dependent on the location model (Pan, 1999). However, this problem can be easily eliminated through replication, and thus this section examines methods for replicated experiments.

Commonly known traditional measures for identifying dispersion effects involve the comparison of the within treatment sample standard deviations (s_i) or the within treatment sample variances (s_i^2) between treatment groups (Box and Meyer, 1986; Nair and Pregibon, 1988). It is also common to take the natural logarithm of s_i or $s_i + 1$ in order to normalize the results for more accurate analysis (Bartlett and Kendall, 1946). In these methods, the observations in each treatment group are reduced to a single summary statistic for further analysis, opposed to using all n observations as would be done for testing location effects. This results in a loss in degrees of freedom (Mackertich et al., 2003).

Taguchi (1986) proposed various signal-to-noise ratios which each aim to achieve a specific outcome. For example the nominal-the-best ratio aims to achieve a

specific target value while maintaining minimum variation. The main difference in this analysis is that it examines both the location and dispersion effects simultaneously as a single metric. Although these measures may be effective for specific quality control outcomes, simulation studies have proven them not to be the most effective in terms of Type I error rates or power when analyzing dispersion effects alone (Mackertich et al., 2003; Dingus, 2005).

It was not until later that Levene-type transformations were considered as possible methods to identify dispersion effects as opposed to simply testing for equality of variances. Mackertich et al. (2003) suggested transforming each individual observation, as Levene did, such that each observation now provides a measure of dispersion. These transformed observations can then be analyzed using an ANOVA. The new transformed response variable contains the same number of observations as the original data, thus maintaining the same degrees of freedom. It was expected that this increase in degrees of freedom would lead to an increase in power compared to the traditional measures (Mackertich et al., 2003).

Levene-type transformations of the response variable involve taking the absolute deviations from a measure of central tendency for each observation within a treatment group. If the original response variable is normally distributed, this transformation will result in a half-normal or folded distribution that has been folded at zero (Leone et al., 1961). This resulting distribution is heavily right skewed and would no longer satisfy the assumption for normality. Although this will not cause a problem with large replication sizes due to the central limit theorem, it may be of concern for small replication sizes, which are common in factorial designs. It is

for this reason that transformations of the absolute deviations were examined in an attempt to normalize the resulting half-normal distribution.

Mackertich et al. (2003) examined two proposed transformations. They examined the absolute deviation from the mean, $|y_{ij} - \bar{y}_i|$, as well as the absolute deviation from the mean raised to the power of 0.42, $|y_{ij} - \bar{y}_i|^{0.42}$. This power transformation is an approximate normalizing transformation and comes from the Kullback-Leibler information, which is a measure of similarity between two distributions. It is expected that this power will most closely approximate a normal distribution. It was determined that both transformations resulted in an inflated Type I error rate, but did produce increased power when an adjustment for the Type I error rate was conducted (Mackertich et al., 2003).

Similar to Brown and Forsythe's extension of the Levene test, it would seem that other measures of central tendency could be used in this context as well. In a PhD dissertation by Dingus (2005), an extensive comparison of thirty-seven different dispersion measures is conducted. These measures include variations of the traditional measures, Taguchi's signal-to-noise ratios, absolute residuals, absolute deviations from the mean and absolute deviations from the median.

After many simulations and two phases of the study, Dingus (2005) reported that the natural logarithm of the absolute deviation from the mean, $\ln|y_{ij} - \bar{y}_i|$, produces significance levels close to the nominal level and the highest power for data following a normal distribution, however it does not perform well under departures from normality. The natural logarithm of the absolute deviation from the median trimmed by the minimum value, $\ln|y_{ij} - \tilde{y}_i|_{-1}$, where $\tilde{y}_i$ is the within treatment me-

dian, is both a robust measure and produces the highest power under non-normality, specifically for the Cauchy(0,1) and exponential(1) distributions (Dingus, 2005).

Chapter 3

Proposed Simplified Method

A simplified version of the method which Dingus (2005) proposed in her dissertation will be invetigated in this thesis. Dingus (2005) uses the following test statistic from Scheffé (1959) on the transformed response variable

$$M = \frac{(\mathbf{A}\hat{\gamma})'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\gamma})}{(f(\mathbf{Y}) - \mathbf{X}\hat{\gamma})'(f(\mathbf{Y}) - \mathbf{X}\hat{\gamma})} \tag{3.1}$$

to test the hypothesis

$$H_0 : \gamma_4 = 0 \tag{3.2}$$

where $f(\mathbf{Y})$ is a vector of the transformed response variable and $\mathbf{X}$ is the design matrix for a factorial design, in this case a 2^{5-1} fractional factorial containing 5 main effects and all two-way interactions, resulting in a matrix with 16 columns and N rows. $\mathbf{A} = [\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ ]'$ and is an indicator variable of length $2^{5-1} = 16$ corresponding to the variables of interest from the null hypothesis. γ is a vector of dispersion effects where γ_0 is the overall variation and the remaining entries are dispersion effects contributed by its corresponding factor or interaction of factors. In this case, γ_4 is the entry of the γ vector corresponding to the dispersion effect of the

fourth factor and is the variable Dingus (2005) chose to vary in her simulation. The power of the test for H_0 as γ_4 is varied is examined. If the assumptions of ANOVA are met, the M test statistic follows an approximate F distribution with $(l_i - 1)$ and $(N - t)$ degrees of freedom, where l_i is the number of levels of factor i, N is the total number of observations and t is the number of treatment combinations (Dingus, 2005).

Mackertich et al. (2003) used critical values from an F distribution to determine significance of the variable of interest, however, it should be obvious that many of the transformed variables violate the assumption of normality and therefore an F test may not be appropriate. Dingus (2005) therefore compared the M test statistic to an empirical critical value determined through simulation. Under the null hypothesis and at an $\alpha = 0.05$ level of significance, 100,000 iterations of the simulation were performed and the M test statistic was calculated. The 95^{th} percentile value of the M test statistics can then be used as the empirical critical value. This was done for each of the thirty-seven dispersion measures such that a unique critical value was used for each test dependent on the measure. This may be a more accurate method theoretically, but in practice it is unrealistic. Therefore, it is of interest to determine if the method by Mackertich et al. (2003) produces comparable results to the method by Dingus (2005) for other measures of dispersion not studied by Mackertich et al. (2003).

To expand on the method of Mackertich et al. (2003), the following six transformations will be examined in this thesis:

- The absolute deviation from the mean, $|y_{ij} - \bar{y}_i|$
- The natural logarithm of the absolute deviation from the mean, $\ln|y_{ij} - \bar{y}_i|$
- The absolute deviation from the median, $|y_{ij} - \tilde{y}_i|$
- The natural logarithm of the absolute deviation from the median (for even replication sizes only), $\ln|y_{ij} - \tilde{y}_i|$
- The absolute deviation from the median trimmed, $|y_{ij} - \tilde{y}_i|_{-1}$
- The natural logarithm of the absolute deviation from the median trimmed, $\ln|y_{ij} - \tilde{y}_i|_{-1}$

The fourth transformation, $\ln|y_{ij} - \tilde{y}_i|$, is done only for even replication sizes since the absolute deviation from the median for odd replication sizes produces values of zero and the natural logarithm of zero cannot be computed. For the fifth and sixth transformations, the absolute deviation from the median trimmed computes the absolute deviations from the within treatment median and then deletes the smallest observation from each treatment group. The logic for this is that the smallest observation will either be a zero if the replication size is odd or a duplicate if the replication size is even. Eliminating any zero values allows for the examination of the natural logarithm of the deviations from the median for odd replication sizes as well.

These transformations were chosen since they are all Levene-type transformations, the original motivation of this thesis. They also include the recommended transformations from both Mackertich et al. (2003) and Dingus (2005). For each of the transformations, an analysis of variance will be conducted and compared to a

critical value from the F distribution with appropriate degrees of freedom. The empirical Type I error rates and empirical power can be calculated by conducting a large number of simulations. These Type I error rates will be examined for validity and empirical power curves based on the F critical value will be compared to the empirical critical value approach by Dingus (2005) to determine if this simplified method produces comparable results.

It is also of interest to look at how the replication size along with the number of factors affects the power. Mackertich et al. (2003) stated as a general rule that using four or more replications resulted in Type I error rates below 0.1 at the $\alpha = 0.05$ level of significance using the absolute deviation from the mean as the transformed response variable, and that none of the measures tested were effective for a replication size of two. The increase in power and convergence to the nominal level of significance as the replication size increases was examined for a 2^3 factorial design and presented in Table 4 of their paper (Mackertich et al., 2003).

What was not analyzed, however, is how the combined effect of the replication size and the number of factors affects the power of the test. Therefore, it is of interest to determine for a given replication size and power how the size of the effect that is detectable changes as we increase the number of factors. Hence, in this thesis, the following two questions will be the primary focus:

- How adequately does the proposed simplified method perform?
- Holding replication size and power level constant, how does the detectable effect size change as the number of factors increases?

Chapter 4

Simulation

4.1 Methods

Before the simulation process can be described, the model for data generation must be specified. The j^{th} observation in the i^{th} group can be generating using the following model

$$Y_{ij} = \mu_i + \epsilon_{ij} \tag{4.1}$$

where

$$\mu = \mathbf{X}'\beta \tag{4.2}$$

and the $\epsilon_{ij} \sim N(0, \sigma_i)$ such that

$$\sigma = \mathbf{X}'\gamma \tag{4.3}$$

for $i = 1, ..., t$ and $j = 1, ..., r$. t is the number of treatment combinations, calculated as 2^f, and these t groups will be referred to as "treatments". The r observations per treatment will be referred to as "replicates". In this model, μ is a vector of treatment means and σ is the corresponding vector of treatment standard deviations, both of length t. $\mathbf{X}$ is the design matrix for a 2^f full factorial design with all interactions, for

$f = 2, \ldots, 5$, where the first column is a column of 1's and the remaining columns are columns of -1's and $+1$'s representing the low and high level setting of the factor or interaction of factors corresponding to that column. $\mathbf{X}$ has t columns and rt rows. β represents a vector of location effects of length t with the first value being the grand mean and remaining values being factor location effect sizes. Similarly, γ represents a vector of dispersion effects of length t with the first value being the overall variation and remaining values being factor dispersion effect sizes. The γ values are chosen such that all elements of σ are positive values. The individual entries of both the β and γ vectors will be represented with subscripts i for $i = 0, \ldots, t-1$.

It should also be noted that the dispersion effect size can be represented as a proportion of the total variation. For example, γ_1/γ_0 represents the proportion of the total variation accounted for by factor 1.

4.1.1 Comparison to Empirical Critical Values Approach, Dingus (2005)

For a comparison of the Dingus (2005) results to the simplified method, a replication of her Phase I simulation was conducted using the F critical values instead of simulated empirical critical values.

A 2^{5-1} fractional factorial design was used with a single replicate of the design matrix $\mathbf{X}$ shown in Table 4.1. In simulation, the design matrix was expanded to include a column of 1's as the first column and all two-way interactions following the X_5 column. Rows were expanded to include the desired number of replications.

X_1	X_2	X_3	X_4	X_5
−1	−1	−1	−1	1
−1	−1	−1	1	−1
−1	−1	1	−1	−1
−1	−1	1	1	1
−1	1	−1	−1	−1
−1	1	−1	1	1
−1	1	1	−1	1
−1	1	1	1	−1
1	−1	−1	−1	−1
1	−1	−1	1	1
1	−1	1	−1	1
1	−1	1	1	−1
1	1	−1	−1	1
1	1	−1	1	−1
1	1	1	−1	−1
1	1	1	1	1

Table 4.1: Single replicate design matrix after Dingus (2005), Table 14.1

Data was generated according to the model mentioned in section 4.1 with β and γ vectors

$$\beta = [100 \quad 10 \quad -5 \quad 7 \quad 0 \quad 0 \quad 0 \quad 0 \quad 5 \quad 0 \quad -4 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \tag{4.4}$$

$$\gamma = [10 \quad 1 \quad 1.5 \quad -1 \quad \gamma_4 \quad 0.75 \quad 0 \quad 0 \quad 0.5 \quad 0 \quad -0.75 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \tag{4.5}$$

corresponding to equations 14.1 and 14.2 respectively in Dingus (2005). The grand mean is equal to 100 and active location effects correspond to factors X_1, X_2, X_3, X_1X_4 and X_2X_3. The overall variation is equal to 10 and active dispersion effects correspond to factors X_1, X_2, X_3, X_4 (when $\gamma_4 \neq 0$), X_5, X_1X_4, X_2X_3. γ_4 is varied from 0 to 4 in increments of 0.1. Although in Dingus (2005) increments of 0.05 were used, increments of 0.1 should be adequate to compare power while saving computing

time.

The μ and σ vectors are calculated according to equations 4.2 and 4.3, respectively. Data is generated from a standard normal distribution using R statistical software and then rescaled by multiplying by σ and adding μ (R Core Team, 2014). For this comparison, results for only two transformations, $|y_{ij} - \bar{y}_i|$ and $|y_{ij} - \tilde{y}_i|$, are conducted, as these are the only two metrics in common between Dingus (2005) and this thesis. Therefore, only these two transformations will be conducted to create two new response variables. For both response variables, a linear model including all five main effects and all two-way interactions is fit and an analysis of variance is performed. The test statistic is compared to a critical value from an F distribution with 1 and $(16r - 16)$ degrees of freedom, where r is the number of replicates, to test the null hypothesis $H_0 : \gamma_4 = 0$. Whether or not the null hypothesis is rejected is recorded and this process is repeated for 10,000 simulations.

A measure of power, defined as the probability of correctly rejecting a false null hypothesis, can be estimated by counting the number of rejections and dividing by the total number of simulations as follows

$$\text{Empirical Power} = \frac{\text{Number of Times H}_0 \text{ is Rejected}}{10,000} \tag{4.6}$$

In the case that γ_4 is equal to zero and the null hypothesis is true, this is instead a measure of the empirical significance level or Type I error rate, defined as the probability of incorrectly rejecting a true null hypothesis.

The simulation will be done for replication sizes of $r = 4$ and $r = 7$ since

these are the replication sizes reported by Dingus (2005).

4.1.2 Detectable Effect Size

For replication size analysis, two stages were conducted. First, a generic exploratory simulation was performed. Data was again generated following the previously mentioned model. For this simulation, since the location effects are not of interest, the β vector was set as a null vector and the γ vector is defined as

$$\gamma = [10 \quad \gamma_1 \quad 0 \quad 0 \quad \cdots \quad 0] \tag{4.7}$$

such that β and γ are of length t. t is calculated as 2^f, where f is the number of factors in the design. All values except γ_0 and γ_1 are zero. γ_1 is varied from 0.001 to 7 in increments of 0.5.

The μ and σ vectors are calculated according to equations 4.2 and 4.3 respectively. Data is generated from a standard normal distribution using R statistical software and then rescaled by multiplying by σ and adding μ, in this case zero (R Core Team, 2014). The six transformations mentioned in the proposed simplified method chapter are then applied to the response variable to form six new response variables. For the first four transformed response variables ($|y_{ij} - \bar{y}_i|$, $\ln|y_{ij} - \bar{y}_i|$, $|y_{ij} - \tilde{y}_i|$ and $\ln|y_{ij} - \tilde{y}_i|$), a linear model including all main effects and all possible interactions is fit and an analysis of variance is performed. The test statistic is compared to the critical value from an F distribution with 1 and $(tr - t)$ degrees of freedom to test the null hypothesis $H_0 : \gamma_1 = 0$. For the last two transformed response variables

($|y_{ij} - \tilde{y}_i|_{-1}$ and $\ln|y_{ij} - \tilde{y}_i|_{-1}$), a linear model including all main effects and all possible interactions is fit and an analysis of variance is performed. The test statistic is compared to the critical value from an F distribution with 1 and $(tr - 2t)$ degrees of freedom to test the null hypothesis $H_0 : \gamma_1 = 0$. Whether or not the null hypothesis is rejected is recorded and this is repeated through 10,000 simulations. Empirical power is calculated as stated in equation 4.6. The simulation is run for the following replication sizes

$$r = 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 15, 16, 19, 20, 25, 30, 39, 40 \tag{4.8}$$

These replication sizes were chosen in order to provide an equal and comparable number of trials with both even and odd replication sizes. The natural logarithm of the absolute deviation from the median, $\ln|y_{ij} - \tilde{y}_i|$, is conducted only for even replication sizes due to there being a zero value in each of the treatment groups for odd replication sizes.

After the preliminary analysis, additional datapoints are beneficial to produce a better representation of the empirical power curve. In order to compare the detectable effect size between designs with varying numbers of factors and replication sizes, a fixed power level can be chosen. In this thesis, an arbitrary power of 0.8 is used. Simulations for effect sizes at increments of 0.01 are conducted for replication sizes of $r = 4, 7$, and 10 to produce more results around a power of 0.8 for each replication size. The additional effect sizes conducted vary based on the replication size and the number of factors. A power of exactly 0.8 cannot be achieved through the

observed values, so the effect size at a power of 0.8 can be predicted using a model.

A local regression model is fit for each replication and factor combination. One advantage to local regression is that it does not require the specification of a specific function to be fit to the data. Since a power curve follows a non-linear pattern, a non-linear function needs to be fit. Local regression works by fitting a weighted least squares regression model of degree 1, 2 or 3 at a given target point x_0 using only a set of nearby observations. The model is fit by minimizing the squared residuals according to the equation from chapter 7.6 of James et al. (2013) as follows

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2 \tag{4.9}$$

where K_{i0} are weights such that observations that are closer to x_0 have higher weights and observations not included in the set of nearby observations have a weight of zero (James et al., 2013). This equation can be modified depending on the degree of the model fit at each x_0.

The nearby observations are determined by setting a span parameter λ, which can take on values from 0 to 1. As the span is increased, more nearby observations are used and the result will be a smoother curve (Cleveland, 1979). The span is the most important parameter setting of a local regression model and can be determined through a cross-validation procedure (James et al., 2013).

All local regression models are fit using a second-degree polynomial and a span determined through general cross-validation. This is done in R using the 'loess.as()' command from the 'fANCOVA' package (Wang, 2010). This command

automatically chooses a smoothing parameter based on a specified selection criterion. For all models, the selection criterion was set to generalized cross-validation. The predicted effect size and standard error can be determined for a power of 0.8 using the resulting model.

4.2 Results

4.2.1 Comparison Results

The results of the simulation study can be compared to the results in Dingus (2005) by comparing the empirical power curves for a given metric and replication size. A comparison of the absolute deviation from the within treatment mean for a replication size of $r = 4$ can be seen in Figure 4.1. It appears that the two curves are quite similar, although the results from Dingus (2005) hold their Type I error rate at the nominal $\alpha = 0.05$ level of significance while the results from the proposed

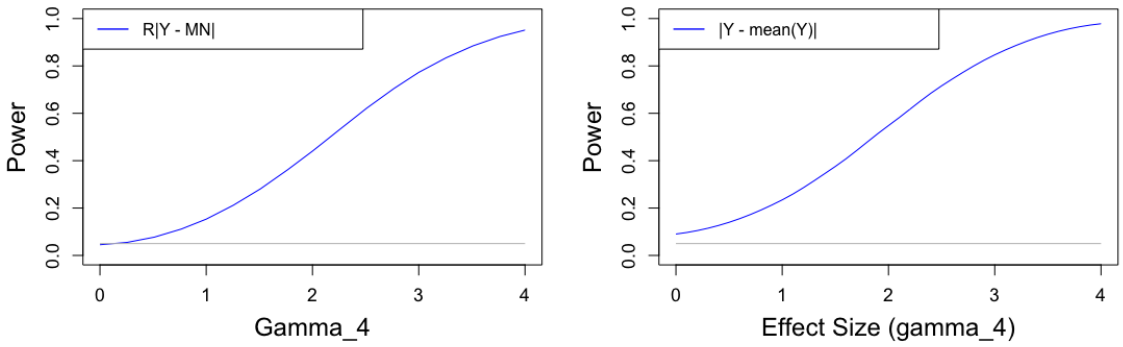


Figure 4.1: Comparison of empirical power curves for the absolute deviation from the mean for replication size of $r = 4$ between the method of Dingus (2005) (left) and proposed simplified method (right). Left figure modified after Dingus (2005), Figure 15.7.

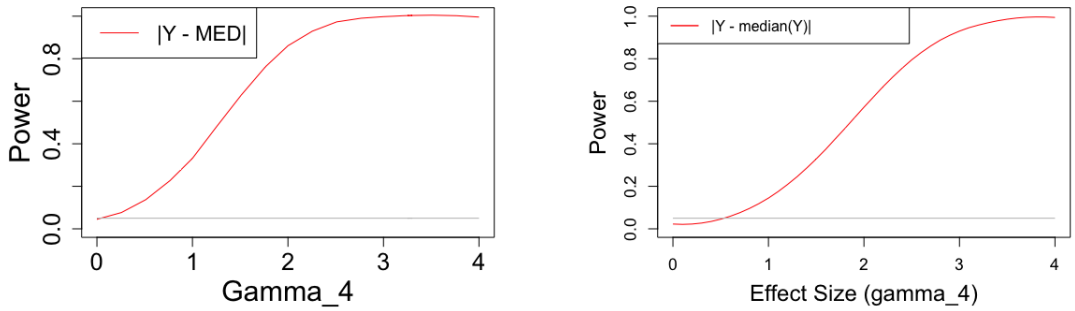


Figure 4.2: Comparison of empirical power curves for the absolute deviation from the median for replication size of $r = 7$ between the method of Dingus (2005) (left) and proposed simplified method (right). Left figure modified after Dingus (2005), Figure 15.14.

simplified method inflate the Type I error rate slightly. This is due to the use of empirical critical values in Dingus (2005) in place of the F critical values. The F critical values should therefore be used with slight caution.

A comparison of the absolute deviation from the within treatment median for a replication size of $r = 7$ can be seen in Figure 4.2. Again, it appears that the two curves are quite similar, although this time the results from the proposed simplified method underestimate the Type I error rate, and thus the power may be slightly more conservative compared to the results from Dingus (2005).

Since both of the metrics produce similar results for the F critical values compared to the empirical critical values in Dingus (2005), they are adequate and will continue to be examined.

The results comparing the various transformations of the response variable can also be examined. The empirical power curves for each of the six transformations for a replication size of $r = 4$ and a factor size of $f = 3$ are presented in Figure 4.3.

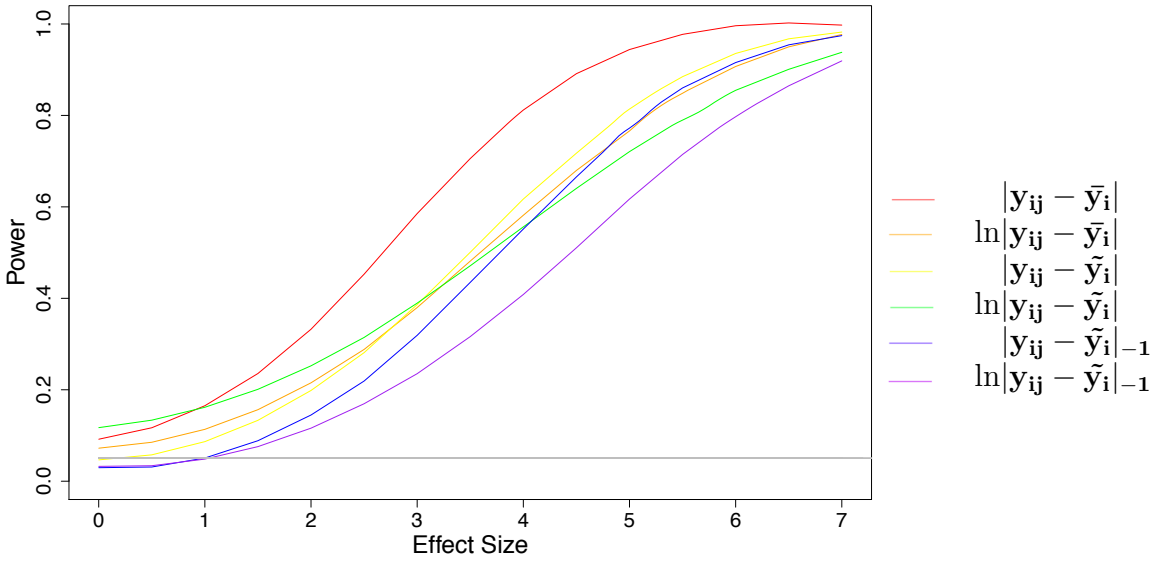


Figure 4.3: Empirical power curves for each of six transformed response variables when replications = 4 and factors = 3

Empirical power curves for additional replication sizes and number of factors can be seen in Figures A.1, A.4, and A.7.

It can be seen that $|y_{ij} - \bar{y}_i|$, $\ln|y_{ij} - \bar{y}_i|$ and $\ln|y_{ij} - \tilde{y}_i|$ all produce increased empirical significance levels and therefore have inflated power curves. Both $|y_{ij} - \tilde{y}_i|_{-1}$ and $\ln|y_{ij} - \tilde{y}_i|_{-1}$ slightly underestimate the significance level and are therefore slightly more conservative than the other options. Of all the transformations, $|y_{ij} - \tilde{y}_i|$ is the only one to maintain the Type I error rate at nominal level of 0.05 and produces the second highest empirical power for an effect size of approximately 3 and larger. Although $|y_{ij} - \bar{y}_i|$ consistently produces a higher empirical power than $|y_{ij} - \tilde{y}_i|$, and both $\ln|y_{ij} - \bar{y}_i|$ and $\ln|y_{ij} - \tilde{y}_i|$ produce higher empirical power for effect sizes less than approximately 3, this may be due to the inflated Type I error rate inflating the empirical power.

As expected, as the number of replications increases, the Type I error rate converges to the nominal level of significance. This can be observed in Figures A.1, A.4 and A.7 in the appendix.

4.2.2 Detectable Effect Size Results

The purpose of this section is to determine how increasing the number of factors in a design while holding the replication size and power constant can change the detectable effect size. This was conducted for replication sizes of $r = 4, 7$, and 10 and power = 0.8 for each of the six transformed response variables. The predicted average effect size $\pm$ the standard error from the fitted local regression model for power = 0.8 for testing H_0: $\gamma_1 = 0$ is presented in Table 4.2.

It can be observed that for all of the transformations, as the number of factors is increased, the effect size of γ_1 detectable with 80% power decreases. This is consistent for all replication sizes examined.

The empirical power curves for each transformation and number of factors when $r = 10$ can also been observed in Figure 4.4. For replication sizes of $r = 4$ and $r = 7$, see Figures A.2 and A.5 respectively. Based on this visual representation, it appears that this decrease in detectable effect size as the number of factors increases is consistent for other levels of power since the curves shift leftward and do not cross.

To estimate the magnitude and test for significance of this apparent decrease in detectable effect size, a model can be fit using the predicted average effect size as the response variable and the number of factors as a continuous explanatory variable. A regression can be fit since the predicted values come from a model fit through

$|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}|$

	2^2	2^3	2^4	2^5
$r = 4$	5.8743 ±0.0679	3.9389 ±0.0503	2.7129 ±0.1044	1.8903 ±0.1460
$r = 7$	4.2989 ±0.0319	2.9396 ±0.1323	2.0519 ±0.2365	1.4379 ±0.1475
$r = 10$	3.1568 ±0.0642	2.4381 ±0.2116	1.7036 ±0.3034	1.1976 ±0.1998

$\ln|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}|$

	2^2	2^3	2^4	2^5
$r = 4$	6.8385 ±0.0625	5.1703 ±0.0312	3.7827 ±0.0513	2.7126 ±0.0524
$r = 7$	5.4491 ±0.0569	4.0124 ±0.0455	2.9002 ±0.1226	2.0596 ±0.0856
$r = 10$	4.6573 ±0.0320	3.4028 ±0.0826	2.4376 ±0.1702	1.7231 ±0.1159

$|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|$

	2^2	2^3	2^4	2^5
$r = 4$	8.3303 ±0.0764	4.9086 ±0.0444	3.2864 ±0.1295	2.2763 ±0.0743
$r = 7$	5.1857 ±0.044	3.7081 ±0.1674	2.4464 ±0.2010	1.7122 ±0.1564
$r = 10$	3.8928 ±0.0526	2.6905 ±0.1579	1.8708 ±0.2946	1.3148 ±0.2358

$\ln|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|$

	2^2	2^3	2^4	2^5
$r = 4$	7.2339 ±0.0880	5.5626 ±0.0450	4.1432 ±0.0330	3.0034 ±0.0326
$r = 10$	5.2537 ±0.0334	3.9074 ±0.0501	2.8077 ±0.1223	2.0102 ±0.1420

$|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|_{-\mathbf{1}}$

	2^2	2^3	2^4	2^5
$r = 4$	8.9249 ±0.0829	5.1464 ±0.0398	3.4067 ±0.1662	2.3500 ±0.0737
$r = 7$	4.7976 ±0.0331	3.2389 ±0.0983	2.2483 ±0.2135	1.5718 ±0.1507
$r = 10$	3.7644 ±0.0656	2.6006 ±0.1416	1.8100 ±0.2556	1.2763 ±0.2300

$\ln|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|_{-\mathbf{1}}$

	2^2	2^3	2^4	2^5
$r = 4$	7.7880 ±0.0900	6.0302 ±0.0411	4.4528 ±0.0430	3.2084 ±0.0336
$r = 7$	5.9197 ±0.0770	4.3338 ±0.0364	3.1423 ±0.1075	2.2512 ±0.1255
$r = 10$	4.8909 ±0.0594	3.5994 ±0.0769	2.5753 ±0.1527	1.8288 ±0.1209

Table 4.2: Predicted average effect sizes ± standard error from fitted local regression model when power = 0.8, for six transformed response variables and replication sizes of 4, 7, and 10

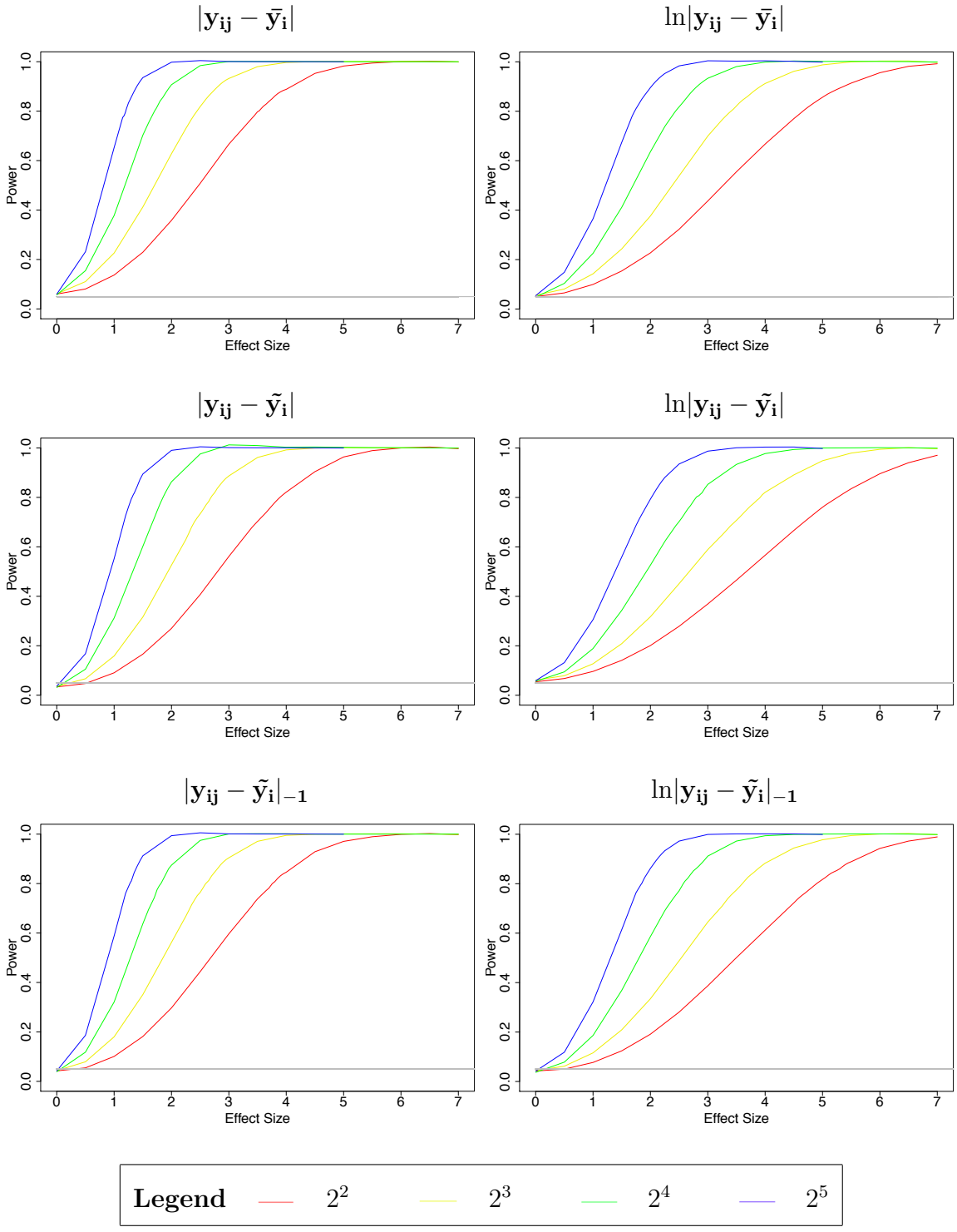


Figure 4.4: Empirical power curves fit using local regression of six transformed response variables by number of factors for replications = 10

randomly generated datapoints, however the data likely violates the assumption of constant variance and the resulting model should be used with caution. By examining a plot of number of factors versus predicted average effect size, it appears that it follows a linear trend as observed in Figure 4.5. Therefore, a linear model is fit for each transformed response.

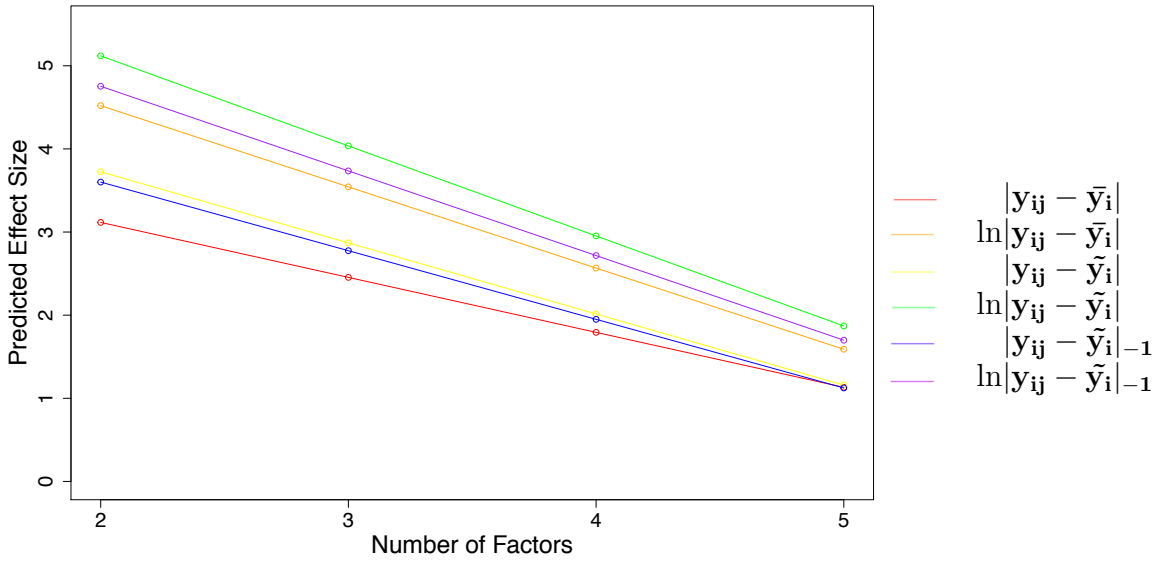


Figure 4.5: Number of factors versus predicted average effect size from the fitted local regression model for each of six transformed response variables, for a replication size of $r = 10$

A fitted linear model of the form $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{y}_i$ is the estimated average effect size for observation i and x_i is the number of factors as a continuous variable for observation i, $i = 1, \ldots, 4$, is fit for each transformed response variable. The estimates of β_1 as well as standard errors, confidence intervals, test statistics and p-values for a replication size of $r = 10$ can be seen in Table 4.3. Linear model plots and β_1 estimates for $r = 4$ and $r = 7$ can be seen in Figures A.3 and A.6, and Tables A.1 and A.2, respectively. All estimates of β_1 are negative, which is consistent with

the results previously observed. The test statistic and p-value for testing H_0: $\beta_1 = 0$ are also presented and are statistically significant at the $\alpha = 0.05$ level of significance for all six of the transformed response variables. It can therefore be concluded that as the number of factors is increased for a fixed replication size, the detectable effect size significantly decreases.

	Estimate	SE	C.I	t-statistic	p-value
$\lvert\mathbf{y_{ij}} - \mathbf{\bar{y}_i}\rvert$	-0.661	0.038	(-0.735 , -0.587)	-17.49	0.0033
$\ln\lvert\mathbf{y_{ij}} - \mathbf{\bar{y}_i}\rvert$	-0.977	0.085	(-1.144 , -0.810)	-11.43	0.0076
$\lvert\mathbf{y_{ij}} - \mathbf{\tilde{y}_i}\rvert$	-0.855	0.103	(-1.057 , -0.653)	-8.34	0.0141
$\ln\lvert\mathbf{y_{ij}} - \mathbf{\tilde{y}_i}\rvert$	-1.083	0.087	(-1.254 , -0.912)	-12.47	0.0064
$\lvert\mathbf{y_{ij}} - \mathbf{\tilde{y}_i}\rvert_{-\mathbf{1}}$	-0.826	0.100	(-1.022 , -0.630)	-8.26	0.0144
$\ln\lvert\mathbf{y_{ij}} - \mathbf{\tilde{y}_i}\rvert_{-\mathbf{1}}$	-1.018	0.088	(-1.190 , -0.846)	-11.60	0.0074

Table 4.3: Coefficient estimates of β_1, standard errors, confidence intervals and test statistics with associated p-values for testing H_0: $\beta_1 = 0$ from fitted linear models for each of six transformed response variables, for a replication size of $r = 10$.

The coefficients of the linear models can also be represented as a decrease in the proportion of the total variation contributed by the factor of interest. Using $\ln|y_{ij} - \tilde{y}_i|$ as an example, when the number of factors increases by 1, the average detectable effect size decreases by 10.83% of the total variation.

Chapter 5

Application

5.1 Butterfat Dataset

The first application uses a dataset retrieved from Hand et al. (1993) *A Handbook of Small Datasets*. This particular dataset is found in Sokal and Rolhf (1981).

The dataset comprises average butterfat percentages of pure-bred dairy cattle as the response variable. The first factor is breed at 5 levels (Ayrshire, Canadian, Guernsey, Holstein-Fresian, and Jersey) and the second factor is age at 2 levels (mature and 2-yr). For each breed-age combination, 10 cows were randomly sampled and their average butterfat percentage measured. Although this is an observational study and not a designed experiment, it can be treated as a $2^1 \times 5^1$ balanced full factorial design.

This dataset is commonly analyzed using an analysis of variance for the location effects of breed and age. From this analysis, it is clear that breed had a significant effect on average butterfat percentage, and conducting a Tukey's Honest Significant Difference test for all pairwise differences shows that all of the two-way breed combinations are significantly different with the exception of the Jersey-Guernsey

combination. Tables of results from location effect analysis can be found in Tables B.1 and B.2.

By looking at boxplots of the untransformed mean butterfat percentages by breed (Figure 5.1) and conducting a Levene's test for homogeneity of variance among breed groups using both the mean and the median as a measure of central tendency (Table 5.1), it can be observed that there likely is a dispersion effect of breed. With

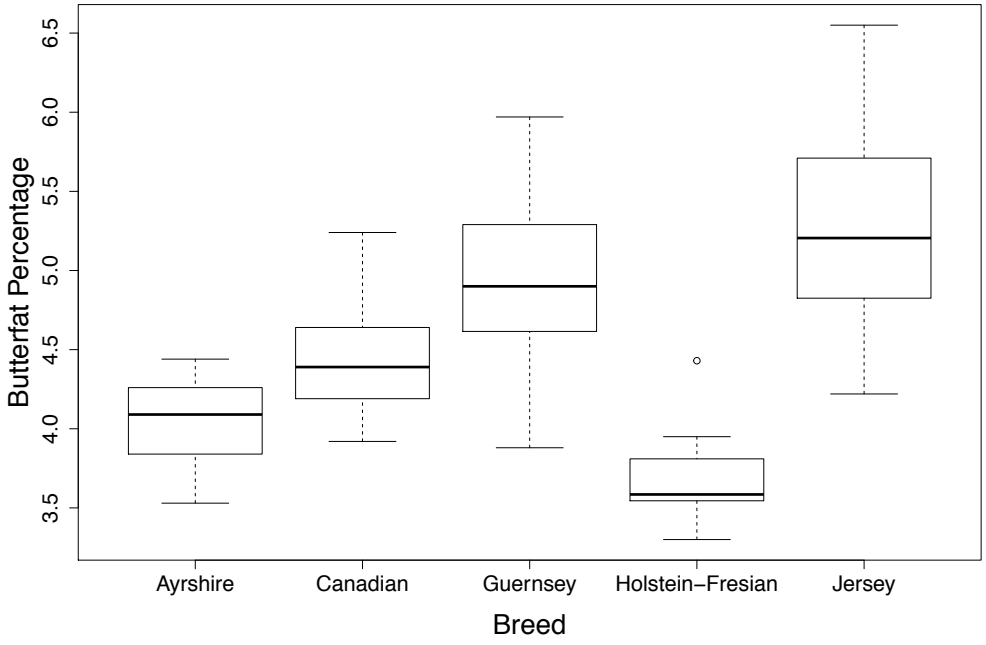


Figure 5.1: Boxplot of butterfat percentage by breed

	Df	F-value	p-value
Mean			
Breed	4	3.77	0.0069
Residual	95		
Median			
Breed	4	3.19	0.0166
Residual	95		

Table 5.1: Levene's test for homogeneity of variance results by breed.

a replication size of 10, the dataset can be analyzed using the six Levene-type transformations in order to identify the dispersion effects.

An analysis of variance was conducted on all of the transformed response variables and results can be seen in Table 5.2. It was consistently found that there was not a significant dispersion effect of age or of the interaction between age and breed. There was a significant dispersion effect of breed for some of the transformations.

$|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}|$

	Df	F-value	p-value
Breed	4	3.74	0.0073
Age	1	0.14	0.7063
Breed×Age	4	0.97	0.4241
Residual	90		

$\ln(|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}| + \mathbf{1})$

	Df	F-value	p-value
Breed	4	1.65	0.1680
Age	1	0.37	0.5444
Breed×Age	4	1.54	0.1984
Residual	90		

$|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|$

	Df	F-value	p-value
Breed	4	3.03	0.0214
Age	1	0.07	0.7849
Breed×Age	4	0.81	0.5221
Residual	90		

$\ln|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|$

	Df	F-value	p-value
Breed	4	1.37	0.2518
Age	1	0.02	0.8888
Breed×Age	4	1.85	0.1253
Residual	90		

$|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|_{-\mathbf{1}}$

	Df	F-value	p-value
Breed	4	3.45	0.0118
Age	1	0.09	0.7615
Breed×Age	4	0.86	0.4918
Residual	80		

$\ln|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|_{-\mathbf{1}}$

	Df	F-value	p-value
Breed	4	1.65	0.1698
Age	1	0.00	0.9490
Breed×Age	4	1.69	0.1605
Residual	80		

Table 5.2: ANOVA results for six transformed response variables for Butterfat dataset.

It can be observed that the dispersion effect of breed is not significant when the natural logarithm of the transformation is taken, but is significant for the absolute value transformations themselves.

For the three transformations that produced a significant breed dispersion

effect, a Tukey's HSD test was performed for all pairwise breed comparisons (Table B.3, B.4 and B.5). All three tests resulted in a significant difference in variation between the Jersey and Holstein-Fresian breeds, and a significant difference in variation was also found between Jersey and Ayrshire breeds when using the absolute deviation from the mean and the absolute deviation from the median trimmed.

5.2 Ceramic Dataset

The second application uses a dataset collected by Said Jahanmir in 1996 at the NIST Ceramics Division and Lisa Gill and James Filliben from the NIST Statistical Engineering Division conducted the study. The data was retrieved from NIST/SEMATECH (2012).

The study was conducted to determine the effect of various machining factors on the strength of ceramic material. The full study contains four factors: table speed, down feed rate, wheel grit size and direction. The dataset that will be worked with only contains observations when direction is set to longitudinal, and therefore it will only include half of the full dataset.

The response variable is a measure of ceramic strength. The first factor is table speed at 2 levels (0.025 and 0.125), the second factor is down feed rate at 2 levels (0.05 and 0.125), and the third factor is wheel grit size at 2 levels (80 and 150). Each treatment combination is used for two batches, each containing fifteen observations. The study is conducted in eight labs, however each lab only conducts two of the sixteen treatment combinations. For this reason, lab will be treated as a

block. Speed, rate, grit and batch will be treated as the factors. This results in a balanced incomplete block design, with a 2^4 factorial structure within blocks.

This dataset can be analyzed to determine which factors have a location effect on the strength of ceramic. From this analysis, it is concluded that lab, batch, speed, batch×speed interaction, and speed×grit interaction are all significant at the $\alpha = 0.05$ level of significance and grit is significant at the $\alpha = 0.1$ level of significance. These results can be seen in Table B.6. However, in a machining scenario, reducing variation is also of interest and since each treatment combination contains fifteen replications, the dispersion effects can also be analyzed.

Conducting a Levene's Test for homogeneity of variance using both the mean and the median as a measure of central tendency concludes that there is no significant difference in variation for any of the factors and results can be seen in Table 5.3.

Batch

	Df	F-value	p-value
Mean	1	0.71	0.4003
Residual	478		
Median	1	0.39	0.5349
Residual	478		

Speed

	Df	F-value	p-value
Mean	1	0.90	0.3431
Residual	478		
Median	1	0.97	0.3243
Residual	478		

Rate

	Df	F-value	p-value
Mean	1	0.95	0.3307
Residual	478		
Median	1	0.70	0.4019
Residual	478		

Grit

	Df	F-value	p-value
Mean	1	0.00	0.9998
Residual	478		
Median	1	0.00	0.9709
Residual	478		

Table 5.3: Levene's test for homogeneity of variance results for Ceramic dataset

An analysis of variance using the six transformed response variables was then performed in order to analyze main dispersion effects as well as interaction dispersion effects. In all six of the analyses, none of the main dispersion effects of batch, speed, rate or grit were significant at the $\alpha = 0.05$ level of significance, which is consistent with the results from the Levene's test.

Again, in all six of the analyses, none of the interaction dispersion effects were significant with the exception of the batch×rate interaction using the natural logarithm of the absolute deviation from the median, $\ln|y_{ij} - \tilde{y}_i|$, which was significant at the $\alpha = 0.05$ level of significance.

Since none of the main or interaction dispersion effects are significant in this scenario, the location effects can be analyzed on their own with little concern for changes in dispersion. However, this should always be done with caution since a lack of evidence against the null hypotheses does not necessarily conclude that there is no change in dispersion.

5.3 Survival Times Dataset

The third application uses a dataset retrieved from Hand et al. (1993) *A Handbook of Small Datasets*. This dataset comes from *An Analysis of Transformations (with discussion)* by Box and Cox (1964). The dataset was originally used by Box and Cox to illustrate how their transformation helped to satisfy the assumptions of ANOVA when heterogeneity of variance occurred, therefore this dataset is known to have changes is variation among treatment combinations.

The data measures survival times, in 10 hour units, of animals after a treatment and poison combination was administered. The first factor is treatment at 4 levels (A, B, C and D) and the second factor is poison at 3 levels (I, II and III), resulting in a 3×4 factorial designed experiment. There are 4 replications per treatment-poison combination.

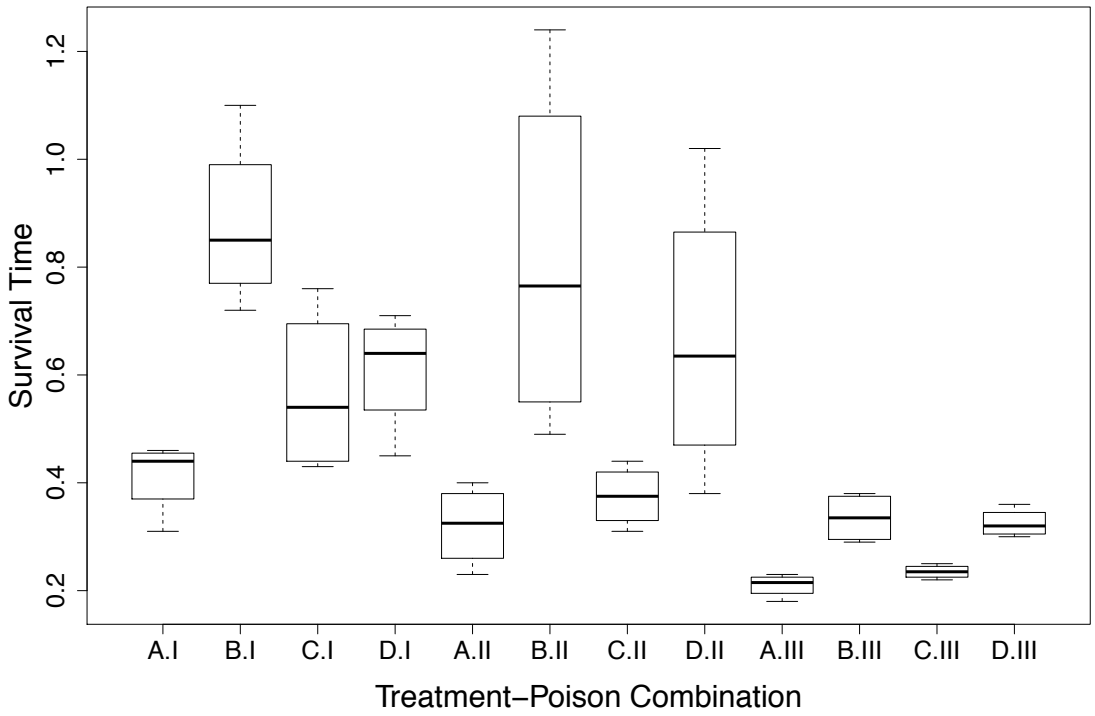


Figure 5.2: Boxplot of mean survival times by Treatment and Poison combination

Examining the boxplot of the survival times by treatment combination in Figure 5.2 makes the heterogeneity of variance very clear. To further validate this expected inequality, a Levene's Test was performed on the treatment factor, the poison factor and the treatment$\times$ poison combination using both the mean and the median as measures of central tendency (Table 5.4). As expected, all of these tests are highly significant and are consistent with the boxplot as well as Box and Cox's

conclusion.

Treatment

	Df	F-value	p-value
Mean	3	6.07	0.0015
Residual	44		
Median	3	5.90	0.0018
Residual	44		

Poison

	Df	F-value	p-value
Mean	2	8.09	0.0010
Residual	45		
Median	2	4.20	0.0213
Residual	45		

Treatment×Poison Combination

	Df	F-value	p-value
Mean	11	4.85	0.0001
Residual	36		
Median	11	4.13	0.0006
Residual	36		

Table 5.4: Levene's test for homogeneity of variance results for Survival Times dataset.

In order to examine which factor levels are driving this change in variation, the six transformations from this thesis were performed and an analysis of variance conducted. However, when the absolute deviation from the within group mean was taken, some values very close to zero arose. The natural logarithm of these deviations was computed and two observations became inadmissible, meaning a model could not be fit. This was resolved by adding a value of 1 to all the absolute deviations from the within group mean before the natural logarithm was taken, $\ln(|y_{ij} - \bar{y_i}| + 1)$.

In all cases, both main dispersion effects of treatment and poison were significant at the $\alpha = 0.05$ level of significance. The interaction term was also significant for three measures at $\alpha = 0.05$, two measures at $\alpha = 0.1$ and not significant for one measure. The results can be seen in Table 5.5.

$|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}|$

	Df	F-value	p-value
Treatment	3	4.32	0.0106
Poison	2	11.60	0.0001
Treatment×Poison	6	2.87	0.0217
Residual	36		

$\ln(|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}| + \mathbf{1})$

	Df	F-value	p-value
Treatment	3	4.40	0.0098
Poison	2	12.44	7.812e-05
Treatment×Poison	6	2.88	0.0215
Residual	36		

$|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|$

	Df	F-value	p-value
Treatment	3	3.78	0.0187
Poison	2	9.72	0.0004
Treatment×Poison	6	2.45	0.0435
Residual	36		

$\ln|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|$

	Df	F-value	p-value
Treatment	3	6.23	0.0016
Poison	2	13.51	5.267e-07
Treatment×Poison	6	2.31	0.0545
Residual	36		

$|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|_{-\mathbf{1}}$

	Df	F-value	p-value
Treatment	3	3.45	0.0324
Poison	2	8.89	0.0013
Treatment×Poison	6	2.08	0.0924
Residual	24		

$\ln|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}|_{-\mathbf{1}}$

	Df	F-value	p-value
Treatment	3	4.12	0.0172
Poison	2	10.45	4.062e-05
Treatment×Poison	6	1.35	0.2741
Residual	24		

Table 5.5: ANOVA results for six transformed response variables for Survival Times dataset.

Conducting a Tukey's Honest Significant Difference test produces an agreement between all six measures that there is a significant difference between treatments B and A, poisons III and I, and poisons III and II. The natural logarithm of the absolute deviation from the within group median also produces a significant difference between treatments C and B. The Tukey's HSD results can be seen in Tables B.8 and B.9.

In this case of very large dispersion effects, all six measures appear to perform similarly, however the extreme outliers that arise from certain measures are of concern.

Chapter 6

Discussion

In response to the first question posed in this thesis as to how adequately the proposed simplified method performs, we can conclude that it does have validity, but with limitations. It is apparent that the use of the F critical values compared to the empirical critical values produce an increase in the Type I error rate, especially when the replication size is small. As the replication and/or number of factors increases, this Type I error rate converges to the nominal significance level, as expected. However, the inflation of the Type I error rate was not as severe for certain measures. The absolute deviation from the within treatment median, $|y_{ij} - \tilde{y}_i|$, produced Type I error rates close to the nominal significance level and resulted in a higher empirical power compared to other measures, even for small replication sizes. This leads us to believe that this measure is both a robust and powerful transformation for detecting dispersion effects.

When addressing the second research question that examines what happens when increasing the number of factors, the decrease in detectable effect size while holding replication size and power level constant was quantified. As the number of factors is increased, the test is able to detect smaller changes in the overall variation that are contributed by a specific factor. From the slope estimates for a replication

size of $r = 10$, for five of the six transformations, as the number of factors increases by 1, the detectable effect size decreases by approximately 10% of the total variation.

This proposed simplified method can be modified for other types of designs, which may be an area of future work. It was also apparent from the Survival Times dataset application that there is a problem with simply taking the natural logarithm of the absolute deviations since extreme outliers and inadmissible values can occur. This may be rectified by adding a constant, but the choice of this constant may be a topic for future discussion. It may also be of interest to revisit other transformations in place of the natural logarithm, such as the square root transformation, to examine how they might increase power or better satisfy the assumptions of the F test.

The results from this thesis extend the work of Mackertich et al. (2003) and Dingus (2005) by varying the number of factors examined in the study. This provides greater information about the relationship between the number of factors in the design, the attainable power and the detectable effect sizes and will allow for improved planning of experiments.

Bibliography

Bartlett, M. S. and Kendall, D. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society.* **8**, 128–138.

Boos, D. D. and Brownie, C. (2004). Comparing variances and other measures of dispersion. *Statistical Science.* **19**, 571–578.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological).* **26**, 211–252.

Box, G. E. and Meyer, R. D. (1986). Dispersion effects from fractional designs. *Technometrics.* **28**, 19–27.

Brenneman, W. A. and Nair, V. N. (2001). Methods for identifying dispersion effects in unreplicated factorial experiments: a critical analysis and proposed strategies. *Technometrics.* **43**, 388–405.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association.* **69**, 364–367.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association.* **74**, 829–836.

Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics.* **23**, 351–361.

Dingus, C. A. V. (2005). Designs and methods for the identification of active location and dispersion effects, Ph.D thesis. The Ohio State University.

Gastwirth, J. L., Gel, Y. R., and Miao, W. (2009). The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science*, 343–360.

Hand, D. J., Daly, F., McConway, K., Lunn, D., and Ostrowski, E. (1993). *A Handbook of Small Data Sets.* vol. 1, Boca Raton: CRC Press.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer.

Leone, F., Nelson, L., and Nottingham, R. (1961). The folded normal distribution. *Technometrics.* **3**, 543–550.

Levene, H. (1960). Robust tests for equality of variances. *Contributions to Probability and Statistics.* **1**, 278–292.

Mackertich, N. A., Benneyan, J. C., and Kraus, P. D. (2003). Alternate dispersion measures in replicated factorial experiments. *Unpublished Manuscript*. `http://www1.coe.neu.edu/~benneyan/papers/doe_dispersion.pdf`.

Nair, V. N. and Pregibon, D. (1988). Analyzing dispersion effects from replicated factorial experiments. *Technometrics*. **30**, 247–257.

NIST/SEMATECH (2012). e-Handbook of Statistical Methods`http://www.itl.nist.gov/div898/handbook/eda/section4/eda42a1.htm`.

Pan, G. (1999). The impact of unidentified location effects on dispersion-effects identification from unreplicated factorial designs. *Technometrics*. **41**, 313–326.

Parra-Frutos, I. (2009). The behaviour of the modified Levenes test when data are not normally distributed. *Computational Statistics*. **24**, 671–693.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scheffé, H. (1959). *The Analysis of Variance*. New York: Wiley.

Schultz, B. B. (1985). Levene's test for relative variation. *Systematic Biology*. **34**, 449–456.

Sokal, R. and Rolhf, F. (1981). *Biometry*. San Francisco: W.H Freeman, 2nd ed.

Taguchi, G. (1986). *Introduction to Quality Engineering: Designing Quality Into Products and Processes*. White Plains, NY: Unipub/Kraus.

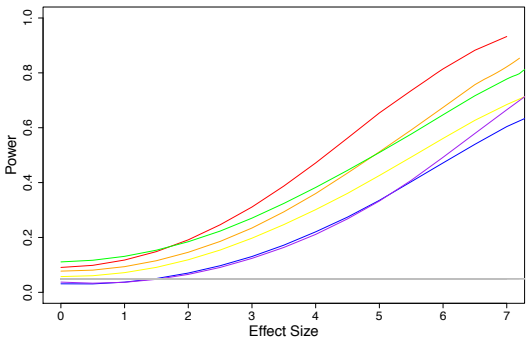
Wang, X.-F. (2010). *fANCOVA: Nonparametric Analysis of Covariance*. R package version 0.5-1.

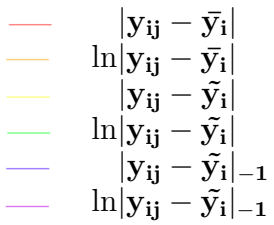
Appendix A

Replication Sizes

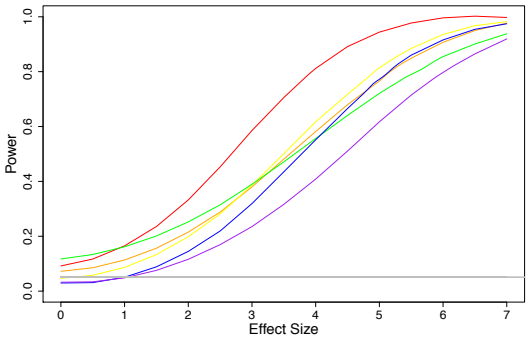
A.1 $r = 4$

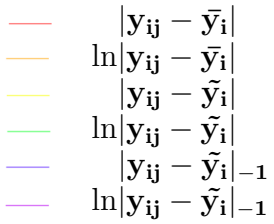
Factors = 2



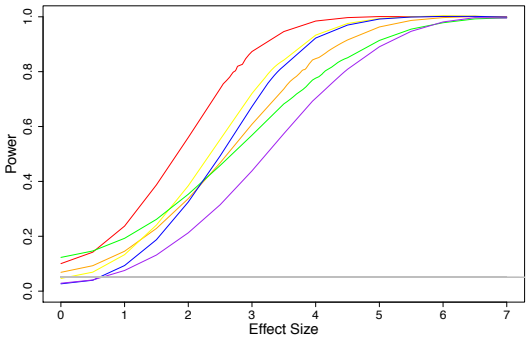


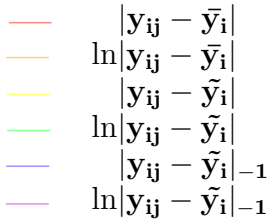
Factors = 3



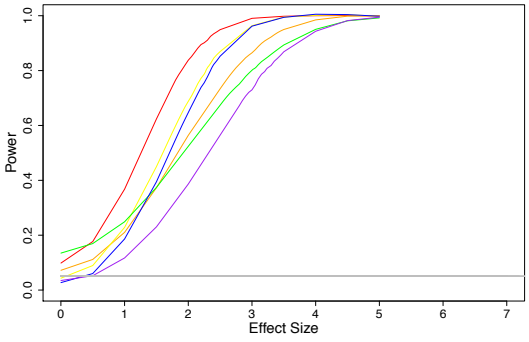


Factors = 4





Factors = 5



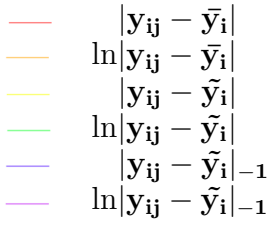


Figure A.1: Empirical power curves for each of six transformed response variables when replications = 4 for $f = 2, \ldots, 5$.

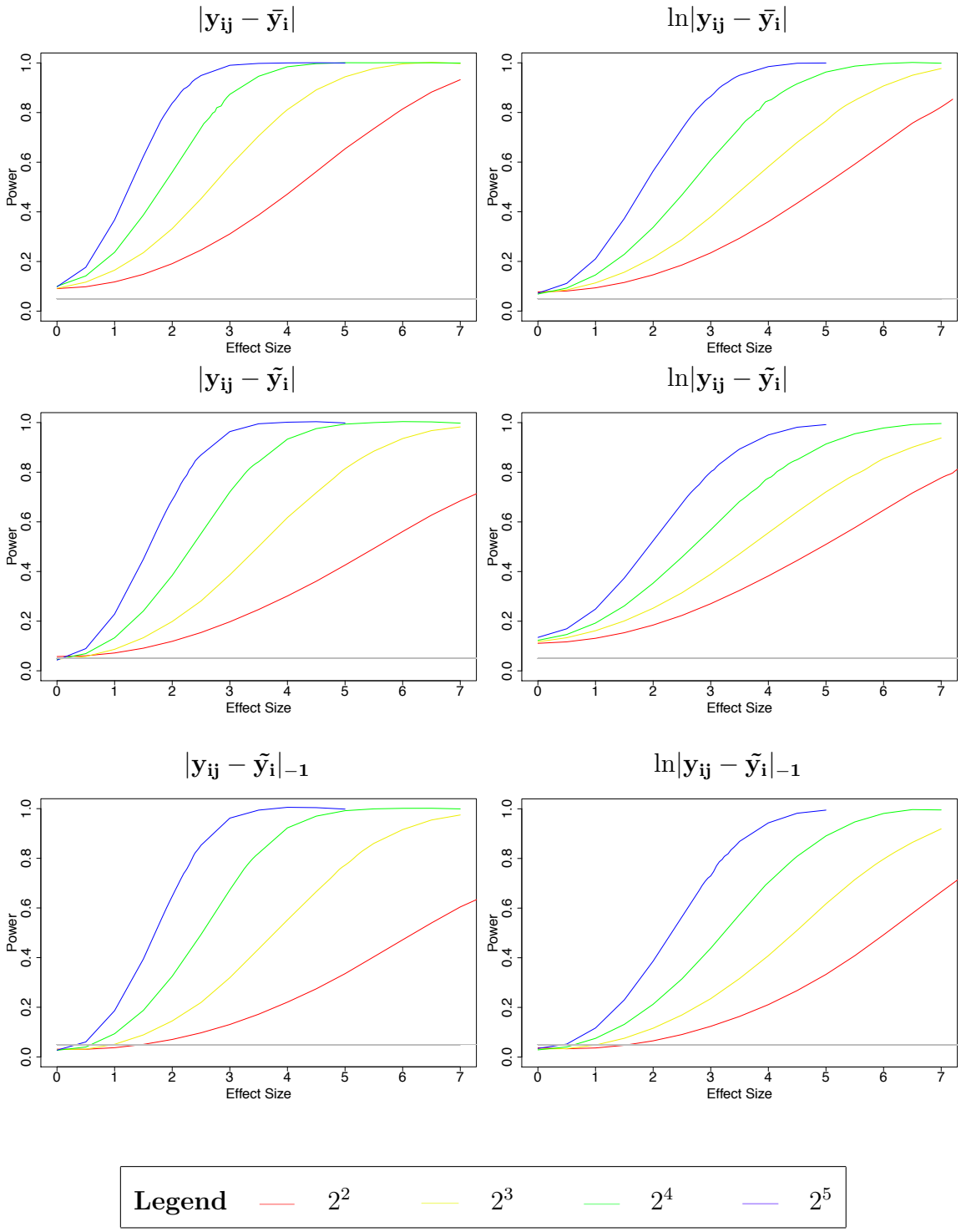


Figure A.2: Empirical power curves fit using local regression of six transformed response variables by number of factors for replications = 4.

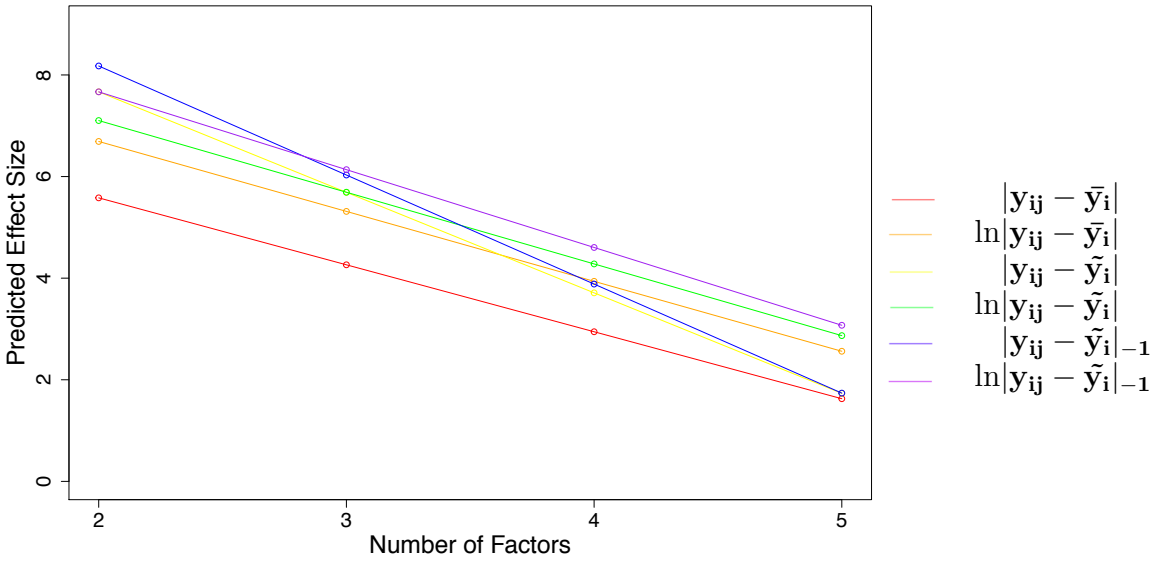


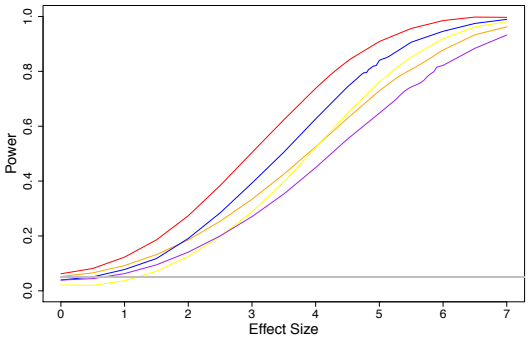
Figure A.3: Number of factors versus predicted average effect size from the fitted local regression model for each of six transformed response variables, for a replication size of $r = 4$.

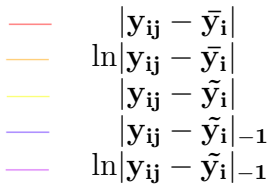
	Estimate	SE	C.I	t-statistic	p-value
$\|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\|$	-1.318	0.177	(-1.665 , -0.971)	-7.43	0.0172
$\ln\|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\|$	-1.377	0.095	(-1.563 , -1.191)	-14.55	0.0047
$\|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\|$	-1.978	0.390	(-2.742 , -1.214)	-5.07	0.0368
$\ln\|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\|$	-1.411	0.084	(-1.576 , -1.246)	-16.79	0.0035
$\|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\|_{-\mathbf{1}}$	-2.146	0.441	(-3.010 , -1.282)	-4.87	0.0397
$\ln\|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\|_{-\mathbf{1}}$	-1.531	0.082	(-1.692 , -1.370)	-18.70	0.0028

Table A.1: Coefficient estimates of β_1, standard errors, confidence intervals and test statistics with associated p-values for testing H_0: $\beta_1 = 0$ from fitted linear models for each of six transformed response variables, for a replication size of $r = 4$.

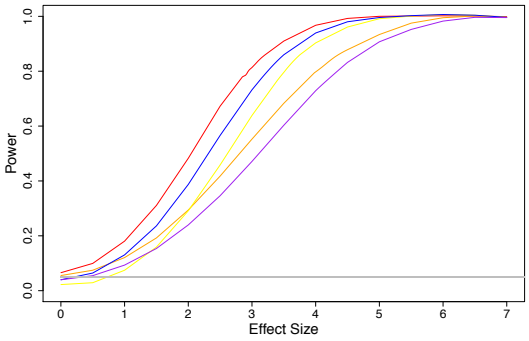
A.2 $r = 7$

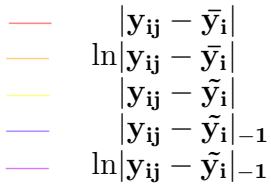
Factors = 2



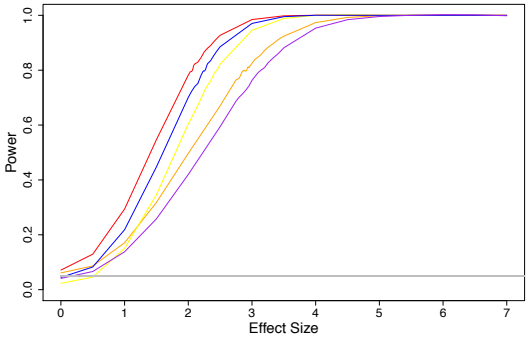


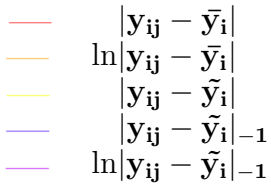
Factors = 3



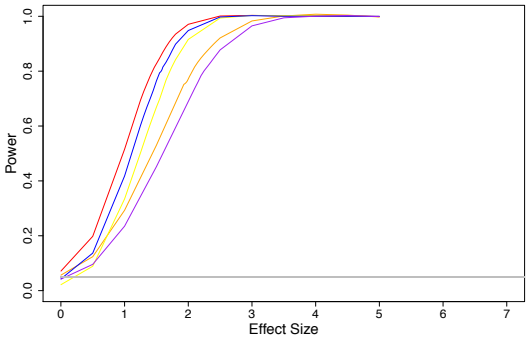


Factors = 4





Factors = 5



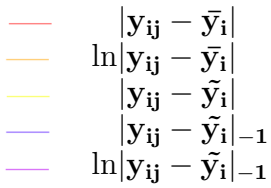


Figure A.4: Empirical power curves for each of six transformed response variables when replications = 7 for $f = 2, \ldots, 5$.

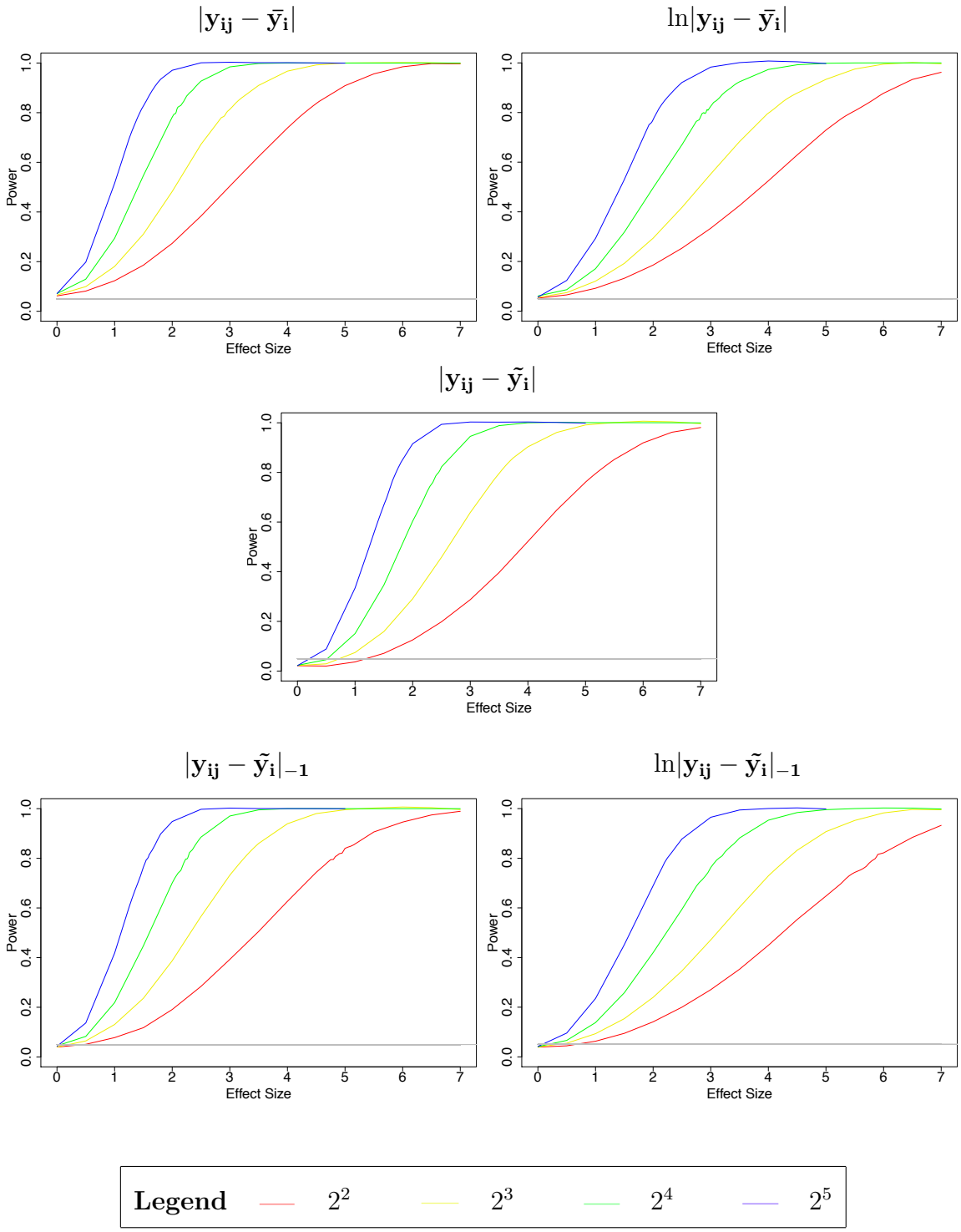


Figure A.5: Empirical power curves fit using local regression of six transformed response variables by number of factors for replications = 7

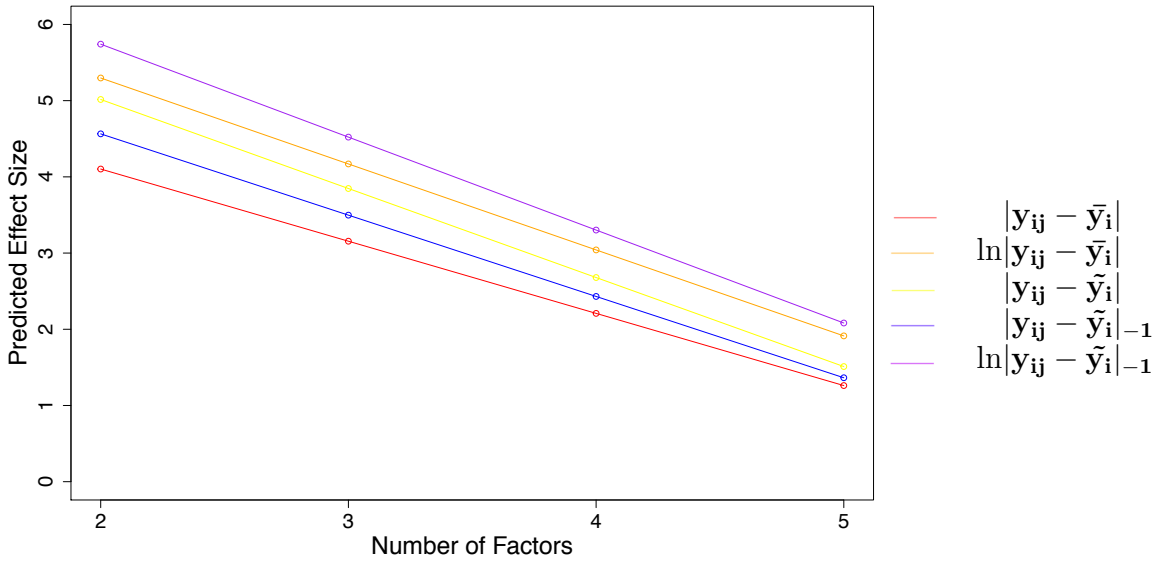


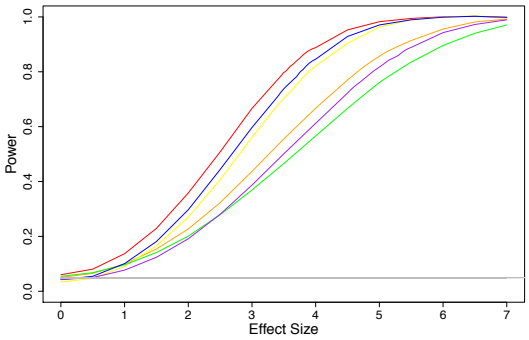
Figure A.6: Number of factors versus predicted average effect size from the fitted local regression model for each of six transformed response variables, for a replication size of $r = 7$.

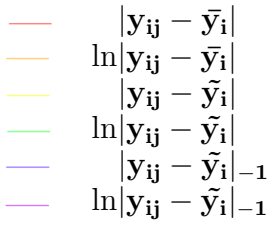
	Estimate	SE	C.I	t-statistic	p-value
$\|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\|$	-0.947	0.119	(-1.180 , -0.714)	-7.98	0.0153
$\ln\|\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\|$	-1.128	0.094	(-1.312 , -0.944)	-11.96	0.0069
$\|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\|$	-1.168	0.120	(-1.403 , -0.933)	-9.77	0.0103
$\|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\|_{-\mathbf{1}}$	-1.067	0.141	(-1.343 , -0.791)	-7.59	0.0169
$\ln\|\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\|_{-\mathbf{1}}$	-1.220	0.110	(-1.436 , -1.004)	-11.08	0.0080

Table A.2: Coefficient estimates of β_1, standard errors, confidence intervals and test statistics with associated p-values for testing H_0: $\beta_1 = 0$ from fitted linear models for each of six transformed response variables, for a replication size of $r = 7$.

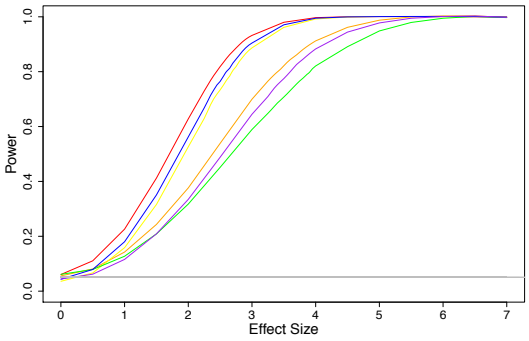
A.3 $r = 10$

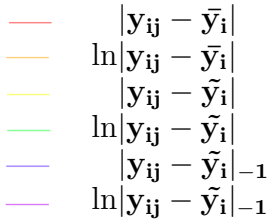
Factors = 2



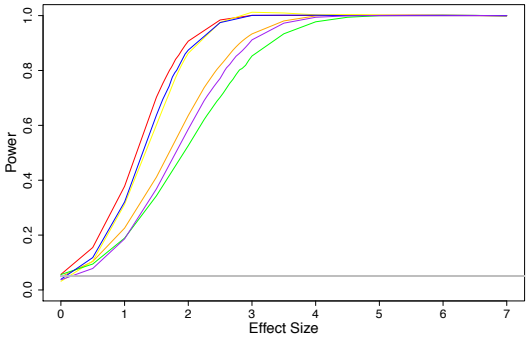


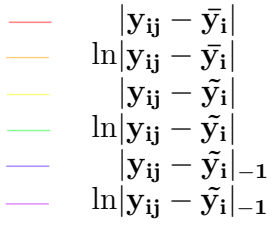
Factors = 3



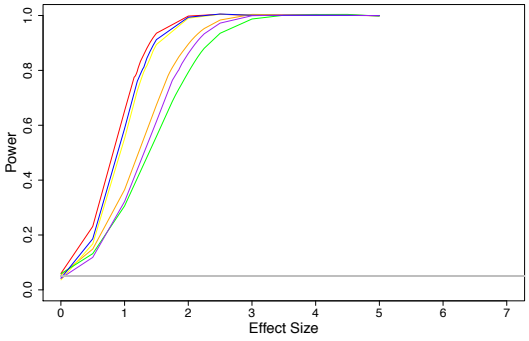


Factors = 4





Factors = 5



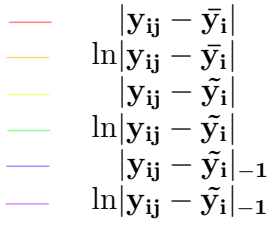


Figure A.7: Empirical power curves for each of six transformed response variables when replications = 10 for $f = 2, \ldots, 5$.

Appendix B

Applications

B.1 Butterfat Dataset

	Df	F-value	p-value
Breed	4	48.59	<2e-16
Age	1	1.17	0.2818
Breed×Age	4	0.38	0.8214
Residual	90		

Table B.1: ANOVA results for mean effects of Butterfat dataset.

	Difference	p-adjusted
Canadian−Ayrshire	0.39	0.0382
Guernsey−Ayrshire	0.89	0.0000
Holstein-Fresian−Ayrshire	-0.39	0.0298
Jersey−Ayrshire	1.23	0.0000
Guernsey−Canadian	0.51	0.0017
Holstein-Fresian−Canadian	-0.77	0.0000
Jersey−Canadian	0.85	0.0000
Holstein-Fresian−Guernsey	-1.28	0.0000
Jersey−Guernsey	0.34	0.0767
Jersey−Holstein-Fresian	1.62	0.0000

Table B.2: Tukey's HSD results for differences in mean location among pairwise breed combinations for Butterfat dataset.

	difference	p-adjusted
Canadian−Ayrshire	0.08	0.8712
Guernsey−Ayrshire	0.17	0.2423
Holstein-Fresian−Ayrshire	-0.02	0.9997
Jersey−Ayrshire	0.24	0.0250
Guernsey−Canadian	0.09	0.8029
Holstein-Fresian−Canadian	-0.09	0.7773
Jersey−Canadian	0.17	0.2390
Holstein-Fresian−Guernsey	-0.18	0.1672
Jersey−Guernsey	0.08	0.8680
Jersey−Holstein-Fresian	0.26	0.0144

Table B.3: Tukey's HSD results for differences in mean dipsersion among pairwise breed combinations for Butterfat dataset using $|y_{ij} - \bar{y}_i|$.

	difference	p-adjusted
Canadian−Ayrshire	0.07	0.9272
Guernsey−Ayrshire	0.16	0.3507
Holstein-Fresian−Ayrshire	-0.02	0.9994
Jersey−Ayrshire	0.23	0.0603
Guernsey−Canadian	0.09	0.8351
Holstein-Fresian−Canadian	-0.09	0.8379
Jersey−Canadian	0.16	0.3255
Holstein-Fresian−Guernsey	-0.18	0.2380
Jersey−Guernsey	0.07	0.9121
Jersey−Holstein-Fresian	0.25	0.0333

Table B.4: Tukey's HSD results for differences in mean dipsersion among pairwise breed combinations for Butterfat dataset using $|y_{ij} - \tilde{y}_i|$.

	difference	p-adjusted
Canadian−Ayrshire	0.08	0.9040
Guernsey−Ayrshire	0.18	0.2832
Holstein-Fresian−Ayrshire	-0.02	0.9994
Jersey−Ayrshire	0.26	0.0374
Guernsey−Canadian	0.10	0.8038
Holstein-Fresian−Canadian	-0.10	0.8054
Jersey−Canadian	0.18	0.2654
Holstein-Fresian−Guernsey	-0.20	0.1876
Jersey−Guernsey	0.08	0.8901
Jersey−Holstein-Fresian	0.28	0.0202

Table B.5: Tukey's HSD results for differences in mean dipsersion among pairwise breed combinations for Butterfat dataset using $|y_{ij} - \tilde{y}_i|_{-1}$.

B.2 Ceramic Dataset

	Df	F-value	p-value
Lab	7	2.66	0.0105
Batch	1	1.17	<2.2e-16
Speed	1	10.75	0.0011
Rate	1	0.43	0.5133
Grit	1	3.79	0.0523
Batch×Speed	1	7.95	0.0050
Batch×Rate	1	2.24	0.1351
Speed×Rate	1	1.16	0.2819
Batch×Grit	1	0.45	0.5039
Speed×Grit	1	3.88	0.0494
Rate×Grit	1	1.97	0.1614
Batch×Speed×Rate	1	0.42	0.5178
Batch×Speed×Grit	1	0.10	0.7530
Batch×Rate×Grit	1	0.06	0.8117
Batch×Speed×Rate×Grit	1	1.30	0.2544
Residual	458		

Table B.6: ANOVA results for mean effects of Ceramic dataset.

B.3 Survival Times Dataset

	B-A	C-A	D-A	C-B	D-B	D-C
$\lvert\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\rvert$	0.0109	0.9166	0.2204	0.0532	0.5396	0.5572
$\ln(\lvert\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\rvert+1)$	0.0097	0.9033	0.2105	0.0525	0.5298	0.5637
$\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert$	0.0173	0.9024	0.2710	0.0862	0.5755	0.6567
$\ln\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert$	0.0011	0.6316	0.1062	0.0274	0.2906	0.6624
$\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert_{-\mathbf{1}}$	0.0354	0.9474	0.2778	0.1094	0.7090	0.5708
$\ln\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert_{-\mathbf{1}}$	0.0153	0.8580	0.2224	0.0860	0.5689	0.6377

Table B.7: Tukey's HSD results for differences in mean dispersion among pairwise treatment combinations using six transformed response variables.

	II-I	III-I	III-II
$\lvert\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\rvert$	0.1333	0.0208	0.0001
$\ln(\lvert\mathbf{y_{ij}} - \bar{\mathbf{y}}_\mathbf{i}\rvert+1)$	0.1584	0.0114	0.0001
$\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert$	0.1430	0.0481	0.0003
$\ln\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert$	0.1446	0.0002	0.0000
$\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert_{-\mathbf{1}}$	0.2047	0.0563	0.0009
$\ln\lvert\mathbf{y_{ij}} - \tilde{\mathbf{y}}_\mathbf{i}\rvert_{-\mathbf{1}}$	0.3538	0.0014	0.0000

Table B.8: Tukey's HSD results for differences in mean dispersion among pairwise poison combinations using six transformed response variables.