Multiple Comparisons of Treatment vs Control Under Unequal Variances Using Parametric Bootstrap

Sarah Alver[1] and Guoyi Zhang[2]

[1, 2]Department of Mathematics and Statistics
[1, 2]University of New Mexico
[1, 2]Albuquerque, NM, 87131
[1]salver@unm.edu
[2]gzhang@unm.edu

Abstract

In one-way analysis of variance (ANOVA) models, it is sometimes of interest to perform simultaneous multiple comparisons of treatment groups with a control group. Dunnett's test is used to test such differences. The assumptions of ANOVA and of Dunnett's test require that the variance of the outcome of interest is the same for each group. However, this assumption is not always met in practice even after transformation. In this research, we developed a parametric bootstrap method for comparing multiple treatment group means against the control group when the constant variance assumption is violated and data are unbalanced. Simulation studies show that the proposed method outperforms Dunnett's test in controlling the type I error under various settings, particularly when data is with heteroscedastic variance and with unbalanced design. An example is presented to illustrate usage of the proposed method. Key words: Parametric bootstrap, Multiple comparison, Unequal variance, Dunnett's test, Simulations, ANOVA, HeteANOVA.

1 Introduction

Consider a one-way analysis of variance (ANOVA) problem with a treatment groups, where the first group is a control group. Let Y_{ij} be the value of the response variable in the jth trial for the ith factor level, $\mu+\alpha_i$ the mean for the ith factor level, $i = 1, 2, \cdots, a, j = 1, 2, \cdots, n_i$. The one-way ANOVA model is as follows:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{1}$$

where $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma_i^2)$, and $\sum_i \alpha_i = 0$.

One may wish to perform multiple comparisons of the treatment groups with the control group, rather than performing all pairwise comparisons. Under the equal variance assumption, Dunnett's test [1, 2] can be used for such purpose and is frequently used in clinical or

pharmacological studies [3, 4, 5, 6]. Dunnett's test compares $a-1$ pairs (each group with the control group), instead of the $\binom{a}{2}$ pairs involved in all pairwise comparisons. Dunnett's test uses the statistic

$$\frac{|\bar{Y}_1 - \bar{Y}_i|}{\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_i)}} \tag{2}$$

where $\bar{Y}_i$ is the sample mean for group i, $i \neq 1$, with α_1 the parameter associated with the control group, and $\hat{\sigma}^2$ is the pooled variance estimate $\frac{\sum_{i=1}^{a}\sum_{n=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2}{\sum_{i=1}^{a} n_i - a}$.

When the assumption of equal variance is violated and data are unbalanced (hereafter called HeteANOVA problem), the results of Dunnett's test are questionable. Many alternative methods were developed for the classical F-test and multiple comparisons for HeteANOVA problems [7, 8, 9]. Among them, the parametric bootstrap (PB) [7] test is shown to be one of the best for testing equality of factor level means. Recently, Zhang [8, 10] proposed PB multiple comparison tests for one-way and two-way ANOVA, which are shown to be competitive.

Inspired by Dunnett's test and PB tests, in this research, we develop a PB test analogous to Dunnett's test, which performs simultaneous multiple comparisons of the treatment groups with the control for the HeteANOVA problem. This research is organized as follows: Section 2 proposes the methodology and presents the algorithm; Section 3 performs a simulation study; Section 4 gives a real example; and Section 5 gives conclusions and discussion of the research.

2 Proposed PB Test and Algorithm

In this section, we develop a PB method for multiple comparisons of treatment groups with the control group for a HeteANOVA problem, and present an algorithm to implement the test.

2.1 Proposed PB Test

Assume without loss of generality that the mean of $\bar{Y}_i$ is zero for all i. Then $\bar{Y}_i \sim N(0, \sigma_i^2/n_i)$ and the sample variance $S_i^2 \sim \frac{\sigma_i^2}{n-1}\chi^2_{(n_i-1)}$ [11]. These can be approximately simulated by pivot variables $\bar{Y}_{Bi} \sim N(0, s_i^2/n_i)$, or equivalently, $\bar{Y}_{Bi} \sim N(0,1)\sqrt{s_i^2/n_i}$, and $S^2_{Bi} \sim \frac{s_i^2}{n-1}\chi^2_{(n_i-1)}$.

Following the procedure from previous papers [7, 8], consider the test statistic in equation (2). We modify this to include the different group variances:

$$T = \frac{|\bar{Y}_1 - \bar{Y}_i|}{\sqrt{(s_1^2/n_1 + s_i^2/n_i)}} \tag{3}$$

We can replace $\bar{Y}_i$ and s_i^2 in equation (3) with $\bar{Y}_{Bi}$ and S^2_{Bi} to obtain a PB pivot variable:

$$T_{PB} = \frac{|\bar{Y}_{B1} - \bar{Y}_{Bi}|}{\sqrt{(S^2_{B1}/n_1 + S^2_{Bi}/n_i)}} \tag{4}$$

We can then simulate a distribution for the test statistic (3), using (4). With this simulated distribution, we can estimate the p-value or obtain a critical value which can be used to

construct confidence intervals. The procedure is shown in the following algorithm.

2.2 Parametric Bootstrap Algorithm for Comparing Multiple Treatment Groups with Control

Algorithm 1

a. For a given $(n_1, n_2, ..., n_a)$, $(\bar{y}_1, \bar{y}_2, ..., \bar{y}_a)$, and $(s_1^2, s_2^2, ..., s_a^2)$, compute the test statistic T in equation (3) for each group paired with the control group.

b. For $l = 1, ..., L$, generate $\bar{Y}_{Bi} \sim N(0,1)\sqrt{s_i^2/n_i}$, and $S_{Bi}^2 \sim \frac{s_i^2}{n-1}\chi_{(n_i-1)}^2$, $i = 1, ..., a$.

c. For each $i \neq 1$, compute the PB pivot variable T_{PB} as in equation (4).

d. D_l = maximum over i of the results from step c.

(end loop).

D is then a simulated distribution for the test statistic. One can use the $1-\alpha$ quantile of D, D_{crit}, as a critical value for a decision rule (i.e. reject $H_0 : \alpha_1 = \alpha_i$ if the test statistic (3) is larger than D_{crit}) or construct a confidence interval using this critical value: $\bar{Y}_i - \bar{Y}_1 \pm D_{crit}\sqrt{(s_i^2/n_i + s_1^2/n_1)}$. As usual, if a p-value is desired, one can compute the proportion of values of D that are greater than the test statistic in (3).

3 Simulations

To evaluate the performance of the algorithm, we simulated 2500 datasets and compared the rejection rate for both Dunnett's Test, using the `DunnettTest` function in the R package `DescTools` [12] and the PB method (Algorithm 1) with $L = 5000$ bootstrap sample mean and variance vectors. We used $a = 6$ treatment groups including the control, with $\boldsymbol{\sigma_1^2} = (1, 1, 1, 1, 1, 1)$, $\boldsymbol{\sigma_2^2} = (0.1, 0.1, 0.1, 0.5, 0.5, 0.5)$, $\boldsymbol{\sigma_3^2} = (1, 1, 1, 0.5, 0.5, 0.5)$, $\boldsymbol{\sigma_4^2} = (0.1, 0.2, 0.3, 0.4, 0.5, 1)$, $\boldsymbol{\sigma_5^2} = (0.3, 0.9, 0.4, 0.7, 0.5, 1)$, and $\boldsymbol{\sigma_6^2} = (0.01, 0.1, 0.1, 0.1, 0.1, 1)$. The sample size vectors used in the simulations were $\boldsymbol{n}_1 = (5, 5, 5, 5, 5, 5)$, $\boldsymbol{n}_2 = (10, 10, 10, 10, 10, 10)$, $\boldsymbol{n}_3 = (3, 3, 4, 5, 6, 6)$, and $\boldsymbol{n}_4 = (4, 6, 8, 12, 16, 20)$. All calculations, simulations and data analysis were performed using R [13].

Results are shown in Table 1. With the equal variance assumption, both Dunnett's test and the PB test give acceptable results. Additionally, when data are balanced, Dunnett's test performs satisfactorily in most heteroscedastic cases. The exception to this is for $\boldsymbol{\sigma_6^2}$. In this case, the simulated p-value for Dunnett's test is higher than the nominal level even with balanced data. This variance vector includes 0.01 which is small, likely leading to an artificially small pooled variance estimate and thus an artificially large test statistic, so the test rejects more often than the nominal level.

The PB test outperforms Dunnett's test, with simulated p-values close to the nominal level for all simulation settings including unequal variance and unbalanced data. In all heteroscedastic cases except $\boldsymbol{\sigma_3^2}$, the proportion rejected for Dunnett's test is too conservative (less than the nominal level) when the data are unbalanced. In these cases, the smaller variances in the simulations are for groups with smaller sample sizes, and larger variances

for groups with larger sample sizes. For these settings, the pooled variance is artificially large, leading to a test statistic that is artificially small. The opposite is true for $\boldsymbol{\sigma_3^2}$, which assigns smaller variances to larger group sizes, so the pooled variance estimate is too small and the test statistic too large.

Table 1: Simulation Results for Multiple Comparisons of Treatment Group Means vs. Control.
Numbers in the table are simulated p-values. We consider four different sample sizes and six different variance vectors as shown in Section 3, with the two different α levels shown.

		$\boldsymbol{\alpha = 0.05}$		$\boldsymbol{\alpha = 0.1}$	
$\boldsymbol{\sigma_1^2}$		Dunnett	PB	Dunnett	PB
	$\boldsymbol{n_1}$	0.0516	0.0420	0.1044	0.0824
	$\boldsymbol{n_2}$	0.0464	0.0360	0.1080	0.0980
	$\boldsymbol{n_3}$	0.0448	0.0384	0.1052	0.0876
	$\boldsymbol{n_4}$	0.0584	0.0592	0.0996	0.1052
$\boldsymbol{\sigma_2^2}$		Dunnett	PB	Dunnett	PB
	$\boldsymbol{n_1}$	0.0464	0.0444	0.0940	0.0872
	$\boldsymbol{n_2}$	0.0420	0.0460	0.0868	0.1056
	$\boldsymbol{n_3}$	0.0100	0.0364	0.0328	0.0836
	$\boldsymbol{n_4}$	0.0016	0.0540	0.0080	0.0996
$\boldsymbol{\sigma_3^2}$		Dunnett	PB	Dunnett	PB
	$\boldsymbol{n_1}$	0.0704	0.0404	0.1272	0.0788
	$\boldsymbol{n_2}$	0.0752	0.0364	0.1368	0.0976
	$\boldsymbol{n_3}$	0.1000	0.0412	0.1836	0.0956
	$\boldsymbol{n_4}$	0.1408	0.0592	0.2064	0.1068
$\boldsymbol{\sigma_4^2}$		Dunnett	PB	Dunnett	PB
	$\boldsymbol{n_1}$	0.0456	0.0472	0.0780	0.0928
	$\boldsymbol{n_2}$	0.0424	0.0424	0.0744	0.1092
	$\boldsymbol{n_3}$	0.0104	0.0376	0.0336	0.0844
	$\boldsymbol{n_4}$	0.0004	0.0496	0.0020	0.0944
$\boldsymbol{\sigma_5^2}$		Dunnett	PB	Dunnett	PB
	$\boldsymbol{n_1}$	0.0348	0.0436	0.0768	0.0892
	$\boldsymbol{n_2}$	0.0320	0.0412	0.0692	0.1056
	$\boldsymbol{n_3}$	0.0184	0.0372	0.0604	0.0872
	$\boldsymbol{n_4}$	0.0096	0.0556	0.0308	0.1016
$\boldsymbol{\sigma_6^2}$		Dunnett	PB	Dunnett	PB
	$\boldsymbol{n_1}$	0.0996	0.0540	0.1392	0.1060
	$\boldsymbol{n_2}$	0.1000	0.0416	0.1440	0.1044
	$\boldsymbol{n_3}$	0.0308	0.0412	0.0628	0.1056
	$\boldsymbol{n_4}$	0.0016	0.0460	0.0036	0.0884

4 Application

An example of the method is shown by applying it to the data discussed by Sanaman and Lear (data downloaded from website by Winner, University of Florida) [14, 15]. The data concerns iron content, in milligrams per liter, found in various depths of seawater. For this example, we considered surface water, where Depth=0, to be the control group. Summary statistics are shown in Table 2. We use five digits for the display as some of the sample variances are quite small. We can see that the variance for 40 feet is somewhat larger than the others, and the variances at the shallower levels are somewhat smaller than those for the deeper levels.

Table 2: Summary Statistics for Iron Data

Depth	$\bar{y}_i$	s_i^2	s_i	n_i
0	0.04267	0.00001	0.00252	3
10	0.03967	0.00006	0.00757	3
30	0.04533	0.00001	0.00231	3
40	0.10867	0.00169	0.04105	3
50	0.10333	0.00020	0.01401	3
100	0.20520	0.00052	0.02282	5

We fit the one-way ANOVA model and then checked assumptions of normality and constant variance. By the Shapiro-Wilk test for normality using the `shapiro.test` function in R (W = 0.9394, p-value = 0.2334), and examination of a normal plot of the standardized residuals (residual plots shown in the appendix), the normality assumption was satisfied. For checking the constant variance assumption, we examined a plot of the standardized residuals against the fitted values from the ANOVA model (Figure 1 in the appendix, right panel). We also performed the Breusch-Pagan test using the function `bptest` from the R package `lmtest` [16]. The p-value from the Breusch-Pagan test was 0.0596, be-

tween the commonly used alpha levels of 0.05 and 0.1, and the residual-fitted plot did appear to indicate non-constant variances.

Several transformations were attempted to satisfy the non-constant variance assumption: log transformation; Box-Cox transformation using $\lambda = -0.2$; and since the units of measurement mg/L could be considered a proportion, the $\sin^{-1}\sqrt{y_{ij}}$ transformation. The λ value for the Box-Cox transformation was found using the `boxcox` function from the R package `MASS` [17]. While the log transformation and the Box-Cox transformation improved the appearance of the fitted-residual plots (shown in Figure 2 of the appendix), none of these improved the p-value from the Breusch-Pagan test.

We performed Dunnett's test, using the previously mentioned function in R, on both the untransformed data and the Box-Cox transformed data. Of note, the normality assumption was still satisfied after the Box-Cox transformation, with W = 0.9554 and p-value = 0.4556 according to the Shapiro-Wilk test. The Dunnett's tests found a significant difference between the iron content of water from the surface (treated as control) and all depths of 40 feet or greater. We then performed the analogous PB test. This test only found a significant difference between the surface and depths of 50 feet or greater. The differences between means, confidence intervals and p-values are shown in Tables 3 and 4 for Dunnett's test and Table 5 for the PB test.

Table 3: Results from Dunnett's Test, Iron Data

	Diff	Lower CI	Upper CI	p-value
10-0	-0.0030	-0.0508	0.0448	0.9998
30-0	0.0027	-0.0451	0.0504	0.9999
40-0	0.0660	0.0182	0.1138	0.0065
50-0	0.0607	0.0129	0.1084	0.0118
100-0	0.1625	0.1198	0.2053	0.0000

Table 4: Results from Dunnett's Test, Box-Cox Iron Data

	Diff	Lower CI	Upper CI	p-value
10-0	-0.1659	-0.8462	0.5143	0.9275
30-0	0.1140	-0.5663	0.7943	0.9835
40-0	1.5183	0.8380	2.1985	0.0001
50-0	1.5144	0.8341	2.1946	0.0004
100-0	2.5269	1.9185	3.1354	0.0000

Table 5: Results from PB Test, Iron Data

	Diff	Lower CI	Upper CI	p-value
10-0	-0.0030	-0.0315	0.0255	0.9852
30-0	0.0027	-0.0095	0.0149	0.8072
40-0	0.0660	-0.0810	0.2130	0.3210
50-0	0.0607	0.0098	0.1115	0.0290
100-0	0.1625	0.0987	0.2263	0.0024

Recall from Table 2 that the measurements taken at 40 feet have a larger variance than the other depths. Thus, the pooled variance estimate could be too small for this group and lead to an artificially large test statistic in the traditional Dunnett's test. In fact, the mean squared error from the ANOVA model for the untransformed data is 0.0004 and the sample variance of the 40-foot depth measurements is 0.0017. A possible practical issue with these results is that if the goal was to get the most iron-rich water from as shallow depth as possible, knowing that the surface was not rich enough, obtaining the water from 40 feet deep could still yield samples that are not as high in iron as desired. A limitation

of this example is that the sample sizes from each depth are small.

5 Conclusions and Discussion

In this research, we looked at Dunnett's test from a parametric bootstrap view and proposed a PB test for comparing treatment groups with the control. Simulation results show that both Dunnett's test and the PB test give acceptable results under the equal variance assumption. Additionally, when data are balanced, Dunnett's test performs satisfactorily in most heteroscedastic cases. However, for heteANOVA problems when equal variance and balanced data assumptions are both violated, Dunnett's test no longer provides reasonable nominal levels, while the proposed PB method works well. From the example, we see that the classical way of transformation to deal with unequal variance is not guaranteed and interpretation of the results after transformation is difficult. The proposed PB test is robust to violation of equal variance and balanced design, and it is easy to implement.

While Dunnett's test performed satisfactorily with most balanced data cases in simulations, the rejection rate can be much higher or lower than the nominal level for the heteANOVA problem. One reason for this is that if one group's variance is much smaller than the others, the pooled variance estimate will be too large, leading to an artificially small test statistic. Similarly, if one group's variance is much larger than the others, the pooled variance estimate will be too small, leading to an artificially large test statistic.

Some limitations of the proposed PB method are that it requires the normality assumption, so if a particular dataset violates both assumptions, a transformation may still be needed. Additionally, as described in [18] section 4.3, we may need to exercise caution

when making practical decisions based on differences in means between groups with unequal variances. For example, if a lower value of a response is desired, such as blood pressure, a treatment group with a smaller mean and smaller variance may have a smaller probability of achieving the desired outcome than a treatment group with a larger mean but also larger variance. Thus, additional consideration of implications for the practical issue being studied is warranted. This issue is illustrated in the iron data example. Despite these limitations, the proposed PB test is a viable method for performing multiple comparisons of treatment vs control for the heteANOVA problem.

References

[1] Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.

[2] Michael H. Kutner, Christopher J. Nachtsheim, John Neter, and William Li. *Applied Linear Statistical Models*. McGraw-Hill Irwin, New York, NY, 5th edition, 2005.

[3] Tallarida R.J. and Murray R.B. *Dunnett's Test (Comparison with a Control). In: Manual of Pharmacologic Calculations*. Springer, New York, NY, 1987.

[4] K. Strojek, K. H. Yoon, V. Hruba, M. Elze, A. M. Langkilde, and S. Parikh. Effect of dapagliflozin in patients with type 2 diabetes who have inadequate glycaemic control with glimepiride: a randomized, 24-week, double-blind, placebo-controlled trial. *Diabetes, Obesity and Metabolism*, 13(10):928–938, 2011.

[5] Zeynep B. Kutuk, Esra Ergin, Filiz Y Cakir, and Sevil Gurgan. Effects of in-office bleaching agent combined with different desensitizing agents on enamel. *Journal of Applied Oral Science*, 27, 00 2019.

[6] Wendy S C Cheng, Therese L Murphy, Maree T Smith, W Graham E Cooksley, June W Halliday, and Lawrie W Powell. Dose-dependent pharmacokinetics of caffeine in humans: Relevance as a test of quantitative liver function. *Clinical Pharmacology & Therapeutics*, 47(4):516–524, 1990.

[7] K. Krishnamoorthy and Fei Lu. A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics and Data Analysis*, 51:5731–5742, 08 2007.

[8] Guoyi Zhang. A Parametric Bootstrap Approach for One-Way ANOVA Under Unequal Variances with Unbalanced Data. *Communications in Statistics - Simulation and Computation*, 44(4):827–832, 2015.

[9] Li-Wen Xu, Fang-Qin Yang, Aji'erguli Abula, and Shuang Qin. A parametric bootstrap approach for two-way ANOVA in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis*, 115:172–180, 2013.

[10] Guoyi Zhang. Simultaneous confidence intervals for pairwise multiple comparisons in a two-way unbalanced design with unequal variances. *Journal of Statistical Computation and Simulation*, 85(13):2727–2735, 2015.

[11] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, second edition, 2002.

[12] Andri Signorell et mult. al. *DescTools: Tools for Descriptive Statistics*, 2020. R package version 0.99.38.

[13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

[14] Michael Sananman and Donald W. Lear. Iron in chesapeake bay waters. *Chesapeake Science*, 2(3/4):207–209, 1961.

[15] Larry Winner, University of Florida, Dept of Statistics. Miscellaneous Datasets: One-Way ANOVA/Independent Samples t-test. `http://users.stat.ufl.edu/~winner/data/ironwater.dat`, accessed 11/29/2020.

[16] Achim Zeileis and Torsten Hothorn. Diagnostic checking in regression relationships. *R News*, 2(3):7–10, 2002.

[17] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.

[18] Ronald Christensen. *Analysis of Variance, Design, and Regression: Linear Modeling for Unbalanced Data*. CRC Press, Boca Raton, FL, 2nd edition, 2016.

Appendix

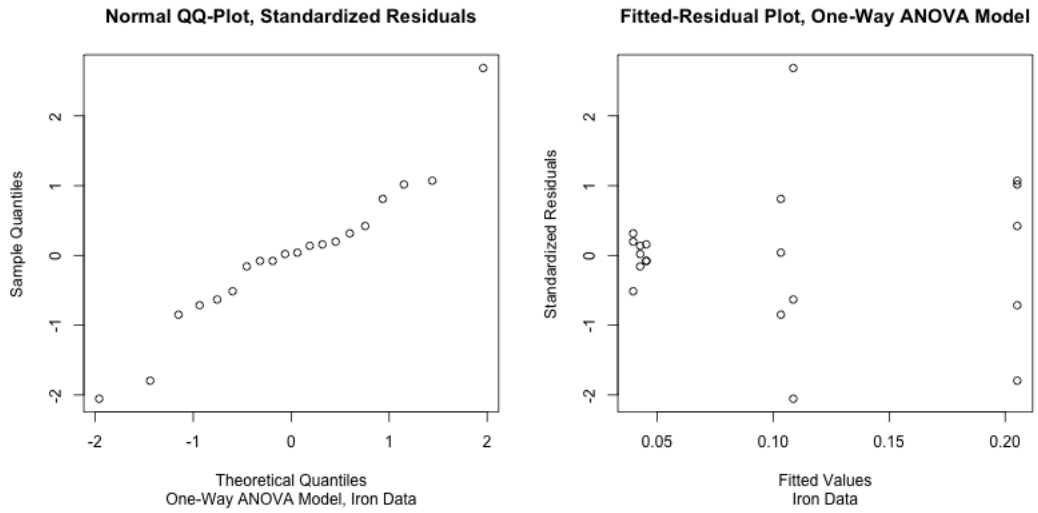

Figure 1: Verification of Assumptions, Iron Data

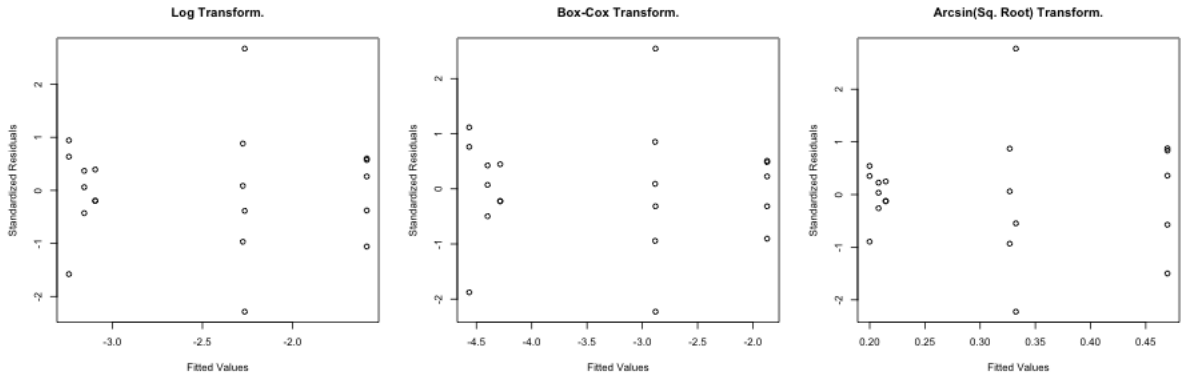

Figure 2: Fitted-Residual Plots after Transformations, Iron Data

R Code: The following code is one way to program the PB test (Algorithm) to simulate a distribution for the PB test statistic. The output here is the test statistic and the p-value, but could be modified to return other values, for example, D_{crit} or confidence intervals.

```
dunnett.PB <- function(L, ns, means, s2){
        D <- rep(0, L)
        r <- length(ns) #number of groups
        pairs.data <- rep(0, r)
        for(j in 1:r){
                pairs.data[j] <- abs(means[1]-means[j])/ #the first will be 0
                sqrt( (s2[1]/ns[1]) + (s2[j]/ns[j]))
                        }
        test.stat <- max(pairs.data)
        ##storage vector for diff in group means for bootstrap data:
        pairs <- rep(0, r)
        for(i in 1:L){
                y.B <- rep(0, r)
                s2.B <- rep(0, r)
                for (j in 1:r){
                y.B[j]<- rnorm(1)*sqrt(s2[j]/ns[j])
                s2.B[j] <- rchisq(1, df=(ns[j]-1))*s2[j]/(ns[j]-1)
                pairs[j] <- abs(y.B[1]-y.B[j])/
                sqrt( (s2.B[1]/ns[1]) + (s2.B[j]/ns[j]))
                                }
```

```
            D[i]<- max(pairs)
                    }
        pval <- length(which(D>test.stat))/L
        return(data.frame(test.stat=test.stat, pval=pval))
}
```