Ordinary Least Squares versus Weighted Least Squares: A Monte Carlo Study

John Fraas, Ashland University
Isadore Newman, University of Akron
John Williams, University of North Dakota
Ordinary Least Squares versus
Weighted Least Squares:
A Monte Carlo Study

Isadore Newman
Professor of Education
Coordinator of The Office of Educational Research
The University of Akron

John Fraas
Dean of the School of Arts and Humanities
Professor of Economics
Ashland College

John Williams
Professor of Education
University of North Dakota

Presented at the American Educational Research Association Annual Meeting in San Francisco,
California, 1989

We thank Carol Wessinger for her assistance in coding the data and procuring the regression analyses.
We also thank Dr. Richard L. Einsprr, Assistant Professor of Mathematics, The University of Akron, for verifying the requested weighted least square analyses in SAS.
Introduction

The weighted least squares (WLS) regression procedure has received considerable attention among researchers. A number of those researchers have indicated that the WLS procedure may be the most desirable and effective method for handling violations of some of the underlying assumptions of the ordinary least squares (OLS) solution. For example, the WLS procedure is frequently recommended when the assumption of homogeneity of variance has been violated. In addition, Chatterjee and Price (1977) suggested that when the underlying assumptions for regression are violated and the violation is due to outliers, then a weighted least squares solution may be an effective procedure to use.

Research by Box (1953, 1954), Glass and Hopkins (1984), Glass, Peckham and Sanders (1972) indicates that the OLS procedure is highly robust and unaffected by the violation of the assumption of homogeneity of variance. Wilcox (1984) states that under assumptions of mild homoscedasticity the use of the OLS procedure presents few problems. Under more extreme violations, however, the use of OLS is questionable (Huynh, 1982).
Researchers must be cognizant of the affects that violations of the underlying assumptions have on the power of the statistical test as well as on Type I error rates. Results of studies and points of view presented by Blair and Higgins (1985), Siegel (1956) and Harwell (1988) suggest that when the assumptions of parametric tests, such as OLS procedures, are violated, power of the statistical tests can be improved substantially with the use of rank ordered data transformations. In some situations the improvement in power can be as much as 20 points.

The purpose of this paper is to investigate through a Monte Carlo study the impact that heterogeneity of variance, outliers, and unequal sample sizes have on the error rates of hypothesis testing. The methods of analyses investigated were: OLS analysis of raw scores, OLS analysis of rank ordered scores, WLS analysis of raw scores, and WLS analysis of rank ordered scores.

Methodology

In this Monte Carlo study, eight data sets were established with each data set consisting of three populations (see Table 1). In every data set except Data Set V, the three population sizes contained 540 scores. In Data Set V, Population 1, Population 2 and
### Table 1

**Population Characteristics**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Group Means</th>
<th>Group Standard Deviations</th>
<th>Group Sample Size</th>
<th>Number of Outliers</th>
<th>Effect Size (^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>(\mu_1 = 5)</td>
<td>(\sigma_1 = 3.148)</td>
<td>(n_1 = 30)</td>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 5)</td>
<td>(\sigma_2 = 3.18)</td>
<td>(n_2 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 5)</td>
<td>(\sigma_3 = 3.7)</td>
<td>(n_3 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>II</td>
<td>(\mu_1 = 9)</td>
<td>(\sigma_1 = 4.33)</td>
<td>(n_1 = 30)</td>
<td>None</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 9)</td>
<td>(\sigma_2 = 7.61)</td>
<td>(n_2 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 9)</td>
<td>(\sigma_3 = 5.24)</td>
<td>(n_3 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>III</td>
<td>(\mu_1 = 31)</td>
<td>(\sigma_1 = 17.33)</td>
<td>(n_1 = 30)</td>
<td>None</td>
<td>1.15</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 41)</td>
<td>(\sigma_2 = 17.33)</td>
<td>(n_2 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 51)</td>
<td>(\sigma_3 = 17.33)</td>
<td>(n_3 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>IV</td>
<td>(\mu_1 = 13.60)</td>
<td>(\sigma_1 = 4.73)</td>
<td>(n_1 = 30)</td>
<td>None</td>
<td>2.12</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 4.33)</td>
<td>(\sigma_2 = 2.64)</td>
<td>(n_2 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 8.16)</td>
<td>(\sigma_3 = 5.75)</td>
<td>(n_3 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>V</td>
<td>(\mu_1 = 10.4)</td>
<td>(\sigma_1 = 3.52)</td>
<td>(n_1 = 10)</td>
<td>None</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 5.17)</td>
<td>(\sigma_2 = 2.92)</td>
<td>(n_2 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 7.5)</td>
<td>(\sigma_3 = 5.62)</td>
<td>(n_3 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>VI</td>
<td>(\mu_1 = 5.56)</td>
<td>(\sigma_1 = 4.34)</td>
<td>(n_1 = 30)</td>
<td>1 Outlier</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 4.97)</td>
<td>(\sigma_2 = 3.0)</td>
<td>(n_2 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 4.70)</td>
<td>(\sigma_3 = 3.13)</td>
<td>(n_3 = 30)</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>VII</td>
<td>(\mu_1 = 31.67)</td>
<td>(\sigma_1 = 18.72)</td>
<td>(n_1 = 30)</td>
<td>1 Outlier</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 41.0)</td>
<td>(\sigma_2 = 17.60)</td>
<td>(n_2 = 30)</td>
<td>1 Outlier</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 51.0)</td>
<td>(\sigma_3 = 17.60)</td>
<td>(n_3 = 30)</td>
<td>1 Outlier</td>
<td></td>
</tr>
<tr>
<td>VIII</td>
<td>(\mu_1 = 7.34)</td>
<td>(\sigma_1 = 6.7)</td>
<td>(n_1 = 30)</td>
<td>3 Outliers</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>(\mu_2 = 7.20)</td>
<td>(\sigma_2 = 6.4)</td>
<td>(n_2 = 30)</td>
<td>3 Outliers</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(\mu_3 = 7.34)</td>
<td>(\sigma_3 = 6.7)</td>
<td>(n_3 = 30)</td>
<td>3 Outliers</td>
<td></td>
</tr>
</tbody>
</table>

\(^1\)The effect size is the ratio of the difference between the smallest and largest group mean, and the mean of the standard deviations.
Population 3 consisted of 60, 180, and 300 scores, respectively. One hundred samples were randomly selected from each data set. Except for Data Set V, each of the 100 samples consisted of 30 scores selected from each of the three populations, thus, producing a total sample of 90 scores. For Data Set V, each of the 100 samples consisted of 10 scores, 20 scores and 30 scores randomly selected from Population 1, Population 2 and Population 3, respectively. In all cases the sampling was conducted with the replacement.

The eight data sets had the following characteristics:

1. Data Set I had equal means, similar variances, equal sample sizes and no outliers.

2. Data Set II contained equal means, unequal variances, equal sample sizes and no outliers.

3. Data Set III had unequal means (large effect size), equal variances, equal sample sizes and no outliers.

4. Data Set IV possessed unequal means (large effect size), unequal variances, equal sample sizes and no outliers.

5. Data Set V had unequal means (large effect size), unequal variances, unequal sample sizes, and no outliers.
6. Data Set VI contained unequal means (small effect size), unequal variances, equal sample sizes and one outlier in Population I.

7. Data Set VII had unequal means (large effect size), similar variances, equal means and one outlier in each population.

8. Data Set VIII contained similar means (very small effect size), similar variances, equal sample sizes and three outliers in each population.

See Table 1 for a listing of the characteristics of the eight data sets.

Each of the 100 samples selected from each data set was analyzed by four procedures to determine whether the population means differed. The four procedures were:

1. OLS procedure applied to the raw scores
2. OLS procedure applied to the rank ordered scores
3. WLS procedure applied to the raw scores
4. WLS procedure applied to the rank ordered scores

The application of OLS procedures to the raw and ranked scores was straightforward and does not require explanation. Since WLS procedures can vary, however, a
discussion of the WLS procedures used in this study is necessary.

WLS procedures differentially weight the data points. The weighting scheme, however, can vary from one study to another. In this study each data point was divided by its corresponding error term that was obtained from the OLS analysis of the data. Thus, points with the largest OLS error terms received the smallest weight in the WLS procedure. As suggested by Chatterjee (1977), one can simply think of WLS procedures as applying OLS procedures to transformed data. This is analogous to the relationship between Pearson's r and Spearman's rho. That is, Spearman's rho is the calculation of Pearson's r with rank ordered data.

The use of four procedures to analyze the 100 samples for each of the 8 data sets produced 3200 statistical tests. The results of those 3200 statistical tests are presented in the next section.

Results

The results of the statistical analyses, which tested for significant differences among the population means at an alpha level of .05, are contained in Table 2.
Table 2

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Set</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I</td>
<td>II</td>
<td>III</td>
<td>IV</td>
<td>V</td>
<td>VI</td>
<td>VII</td>
</tr>
<tr>
<td>OLS</td>
<td>3</td>
<td>2</td>
<td>65</td>
<td>99</td>
<td>26</td>
<td>4</td>
<td>50</td>
</tr>
<tr>
<td>OLS on Ranks</td>
<td>3</td>
<td>1</td>
<td>51</td>
<td>99</td>
<td>27</td>
<td>5</td>
<td>47</td>
</tr>
<tr>
<td>WLS</td>
<td>3</td>
<td>2</td>
<td>63</td>
<td>99</td>
<td>52</td>
<td>6</td>
<td>52</td>
</tr>
<tr>
<td>WLS on Ranks</td>
<td>5</td>
<td>1</td>
<td>55</td>
<td>99</td>
<td>50</td>
<td>8</td>
<td>51</td>
</tr>
</tbody>
</table>

Each number in the table indicates the number of statistically significant tests found in the 100 statistical test conducted on the differences among the three population means.
The contents of Table 2 reveal the following:

1. Data Set I - Very little difference existed in the number of statistical significant tests among the four procedures. Close to the expected number of 5 significant differences was calculated by each procedure.

2. Data Set II - Nearly identical numbers of statistically significant tests were found with each of the four methods.

3. Data Set III - The OLS and WLS procedures produce higher numbers of statistically significant tests, 65 and 63, respectively, when applied to the raw data rather than the rank ordered data.

4. Data Set IV - With the very high effect size contained in this data set the four procedures produced identical results of 99 statistically significant tests.
5. Data Set V - For a data set with unequal means, variances and sample sizes, the WLS procedure produced nearly twice as many statistically significant tests as did the OLS procedure.

6. Data Set VI - Similar numbers of statistically significant tests were obtained by the four procedures when one outlier was included in Population 1 and the effect size was small.

7. Data Set VII - Comparable numbers of statistically significant tests were found by the four procedures when 1 outlier was included in each population and the effect size was quite large.

8. Data Set VIII - The number of statistically significant tests recorded for the four procedures were similar when each population contained 3 outliers.

In summary, three points are apparent from the analyses. First, the outliers contained in Data Sets VI through VIII had little impact on the number of
statistical tests recorded for each of the four methods. Second, both OLS and WLS procedures were more powerful when applied to the raw scores rather than the rank ordered scores under the conditions of equal variances and sample sizes. Third, the WLS procedure applied to raw scores and rank ordered scores resulted in higher numbers of statistically significant findings than the OLS procedure when analyzing populations with unequal means, variances, and sample sizes.

Discussion

The limitations of this study must be kept in mind when considering its implications. First, the sample sizes used were relatively large. Thus, the results may not apply to situations in which the sample sizes are small. Second, the study examined only a limited number of the possible combinations formed by the factors of equal and unequal means, equal and unequal variances, equal and unequal sample sizes, and the inclusion and exclusion of outliers. Third, the populations of raw data used in the study possessed very little skewness and their kurtosis figures ranged from 1.37 to 2.67. An investigation of other population distributions may produce different results.

The results of this Monte Carlo study would warrant a number of considerations from educational
researchers. First, the OLS procedure appears to be robust when the assumption of homogeneity of variance is violated and the sample sizes are equal. This result concurs with the results of the studies by Box (1957) and Wilcox (1987). Second, when the variances of the populations are unequal and the sample sizes are unequal, the application of the WLS procedure rather than the OLS procedure to raw data and rank ordered data appears to provide higher levels of power. Third, under the limited number of conditions examined in this study, the WLS procedure did not produce results that differed from the OLS procedure when applied to data that contained outliers.

If the OLS procedure produces different results than the WLS procedure when applied to conditions not investigated in this study, researchers would do well to consider not only under what conditions the results differed but also the research design implications of applying each procedure. Consider the position that there are four ways to handle outliers.

In one procedure the researcher can ignore the problem that outliers create and use OLS procedures even though an underlying assumption has been violated. With the types of populations examined on this study, such a decision is possible.
A second procedure that researchers can use applies OLS procedures to the data after the outliers have been discarded. This course of action assumes that the outliers in the data are anomalies and only serve to distort the researcher's view of the population.

A third approach is the use of WLS procedures. In this case the researcher does not want to ignore the outliers, but rather the researcher has decided to decrease the influence of the outliers.

In a fourth method, the data set can be rank ordered before it is analyzed. A possible caution with using this approach is the manner in which the results are interpreted. If significant differences existed between the group means when the transformed data were used, it could be misleading and/or inappropriate to also state that the group means differed for the raw data. Newman et al. (1976) indicated that extending the interpretation to the raw data may produce a Type VI error; that is, an inconsistency between the research question and the question being tested by the statistical model. Therefore, when using any statistical technique, one has to be careful to insure that the statistical technique reflects the question of interest.
References


