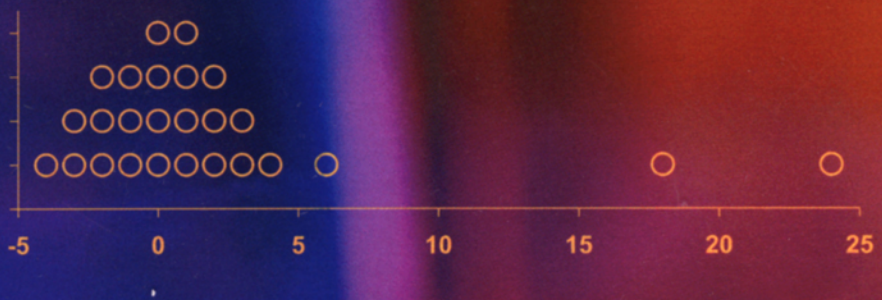
Adaptive Tests of Significance Using Permutations of Residuals with R and SAS®



Thomas W. O'Gorman

WILEY

WWW. LINK AVAILABLE

Adaptive Tests of Significance Using Permutations of Residuals with R and SAS®

Adaptive Tests of Significance Using Permutations of Residuals with R and SAS®

Thomas W. O'Gorman
Northern Illinois University
Division of Statistics
DeKalb, IL

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

O'Gorman, Thomas W.
Adaptive tests of significance using permutations of residuals with R and SAS / Thomas W. O'Gorman.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-92225-5 (hardback)
1. Regression analysis. 2. Computer adaptive testing. 3. R (Computer program language) 4. SAS (Computer file) I. Title.
QA278.2.O35 2012
519.5'36—dc23 2011038049

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To my wife, Martha;
and my children, Kelly
and Tim.

CONTENTS

PREFACE

This book was written to introduce researchers to adaptive tests of significance and to describe the advantages of using these testing methods. Traditional tests of significance, such as the two-sample t test and the t test for slope, are robust in the sense that non-normality of the error distribution often does not dramatically change the level of significance. So, why should we use adaptive tests?

Adaptive tests are used to increase the power when the errors are not normally distributed. In real-world testing situations we rarely know the distribution of the errors, so it is important to know just how the traditional tests compare to adaptive tests with a variety of normal and non-normal error distributions. The power comparisons, which are displayed throughout the book, show that the adaptive test is often much more powerful than the traditional test with many non-normal error distribution.

Adaptive tests use the data to adjust the test procedures. For example, if a researcher wants to perform a two-sample test and the data suggests that the error distributions may be normally distributed, the traditional test procedure is not modified a great deal, so that the resulting test will approximate a two-sample t test. However, if the data contain a few outliers, then the test procedure will be modified to downweight the importance of those outliers.

At first glance these adaptive tests of significance are suspicious. It does not seem right to use the data to modify the test procedure and then use the data again to perform the test. If a test is not properly constructed, it may not maintain its

level of significance; but if a test is properly constructed, it will maintain its level of significance. All of the adaptive tests in this book maintain their significance level because they use permutations methods. These permutation methods have become practical in the last few years because fast computers are now readily available. Most of the adaptive tests in this book can be performed in just a few minutes.

Because permutation methods are used to compute the p-value of adaptive test, some software is necessary. In this book we use R functions and SAS® macros to perform the adaptive weighting and permutation methods. It is not expected that the reader be familiar with both of these languages; either language is sufficient to perform the adaptive tests, confidence intervals, and estimates. Those readers who are not interested in the computational aspects of these tests may choose to skip those sections that describe software.

The basic adaptive testing method is described in the first four chapters. Applications of this method to tests for slope in a regression, tests for main effects and interaction effects in a two-way design, tests for randomized complete block designs, and tests for the analysis of paired data are explained in Chapters 5 through 7. The use of adaptive tests in multicenter clinical trials and in cross-over trials are described in Chapter 8. Chapters 9 and 10 concern multivariate tests and their application to the analysis of repeated measures data. Adaptive confidence intervals, which tend to be narrower than traditional confidence intervals when the errors have a non-normal distribution, are described in Chapter 12.

On a personal note, I became interested in adaptive testing around 1985 while working as a statistical consultant at the University of Iowa. A few years later, while working on variable selection methods in case-control studies with Robert (Skip) Woolson, I became convinced that variables selection methods could be improved if the performance of the testing methods could be improved. After publishing several rank-based adaptive tests, I decided that a more general approach was necessary. In the last ten years I have developed an adaptive test, which is described in this book, that can be used in many testing situations to provide increased power.

I am indebted to many individuals who have contributed to the literature on adaptive tests. The important paper by Hogg, Fisher, and Randles (1975) showed that an adaptive test could maintain its significance level and be relatively powerful. The papers by Büning (1996, 1999, 2002) extended the methods proposed by Hogg. Both Hogg and Büning proposed rank-based adaptive tests that did not use a permutation method.

I am also indebted to Freedman and Lane (1983), who proposed the permutation of residuals method that is used in this book. The papers by Anderson and Legendre (1999) and Anderson and Robinson (2001) increased my understanding of the importance of distinguishing between the various kinds of permutation methods. Most of the ideas concerning permutation tests, which are described in Chapter 4, come from those papers.

All of the simulation studies in this book were programmed in FORTRAN because they could not have been executed in a reasonable amount of time in either R or SAS. In general, programs written in FORTRAN execute much more quickly than those using R code or SAS macros. However, FORTRAN is not usually used for the

analysis of data because it takes much longer to write FORTRAN code than R or SAS code. Hence, in most consulting situations, it is necessary to use either R or SAS to perform the adaptive test. The author would be happy to supply the FORTRAN code for any of the simulations performed in this book to any researcher who is familiar with the use of adjustable dimension arrays in FORTRAN. Please contact the author at *ogorman@math.niu.edu* and describe your interests or concerns to obtain the FORTRAN source code for the simulation study.

I want to acknowledge the support I have been given by my family and colleagues. I would also like to thank Professor Alan Polansky for his encouragement to develop these methods and for his assistance with the R language.

T. W. O'GORMAN

Dekalb, Illinois
November 2011

CHAPTER 1

INTRODUCTION

1.1 WHY USE ADAPTIVE TESTS?

Many adaptive tests have been developed in an effort to improve the performance of tests of significance. We will consider a test of significance to be "adaptive" if the test procedure is modified after the data have been collected and examined. For example, if we are using a certain kind of two-sample adaptive test we would collect the data and calculate selection statistics to determine which two-sample test procedure should be used. If the data appear to be normally distributed, then a Wilcoxon rank-sum test would be used; but if the data appear to contain outliers, then a median test would be used.

Adaptive tests of significance have several advantages over traditional tests. They are usually more powerful than traditional tests when used with linear models having long-tailed or skewed distributions of errors. In addition, they are carefully constructed so that they maintain their level of significance. That is, a properly constructed adaptive test that is designed to maintain a significance level of α will have a probability of rejection of the null hypothesis at or near α when the null hypothesis is true. Hence, adaptive tests are recommended because their statistical properties are often superior to those of traditional tests.

Adaptive Tests of Significance Using Permutations of Residuals with R and SAS®.
By Thomas W. O'Gorman

The adaptive tests described in this book have the following properties:

- The actual level of significance is maintained at or near the nominal significance level of α.
- If the error distribution is long-tailed or skewed, the adaptive test is usually more powerful than the traditional test, sometimes much more powerful.
- If the error distribution is normal, there is little power loss compared to the traditional tests.
- Adaptive test are practical. R code and SAS® macros are available to quickly perform these tests.

The adaptive tests in this book automatically reduce the influence of outliers. They are sometimes said to be robust; but to be clear about robustness, we should describe the two kinds of robustness. A test is said to be robust for size if its actual significance level is quite close to the nominal significance level, even when the usual assumptions are not met. For example, a test that is derived by assuming normality of the error distribution would be robust for size if it maintains its level of significance with non-normal errors. A test is said to be robust for power if it has high power relative to other tests when the usual distributional assumptions are not met. Many traditional tests are robust for size with non-normal errors but are not robust for power. Our objective is to develop adaptive tests that are robust for size and robust for power.

In this chapter we will give a brief history of adaptive tests and will present some of the procedures that are used to develop adaptive tests. We will also show the power advantages of several two-sample adaptive tests. Subsequent chapters will describe the advantages of the adaptive approach for a test of any subset of coefficients in a linear model. In Chapters 3 through 8 we show how adaptive tests can be used in almost all testing situations and that these tests have better properties than the traditional tests. In Chapters 9 and 10 we develop a multivariate adaptive test and show how it can be used to analyze repeated measures data. In Chapter 11 we describe several rank-based approaches to testing and in Chapter 12 we show how adaptive confidence intervals and estimates can be computed. In most chapters, R code and SAS macros will be used to perform the tests.

1.2 A BRIEF HISTORY OF ADAPTIVE TESTS

1.2.1 Early Tests and Estimators

The first two-sample adaptive test that was practical and relatively powerful was proposed by Hogg, Fisher, and Randles (1975). Prior to 1975, the adaptive tests were interesting but not too practical. For example, the test proposed by Hájek (1962) was designed to improve the power by finding scores that would produce a locally most powerful rank test. The test required an estimate of the density function (f) and the first derivative of the function (f'). The problem with this approach is that f and f'

are difficult to estimate unless the samples are very large. Hence, these adaptive tests are not practical and do not appear to be used. See Hogg (1974) for a discussion of the problems associated with this approach.

In order to avoid estimating densities and their derivatives, Hogg (1967) proposed an adaptive procedure that used the sample kurtosis to select one of four estimators of the mean of a symmetric distribution. In that research, four symmetric distributions were considered having various amounts of kurtosis. The idea was to use the selection statistic to select an estimator that would have low variance for samples from that distribution. One difficulty with this approach is that the sample kurtosis is highly variable, so it may sometimes fail to select the correct estimator for that symmetric distribution. In spite of this problem, the robust adaptive estimator had excellent performance with $n = 25$ observations that were generated from the four distributions that were used in that study. In arguing for greater use of these robust methods, Hogg (1967) stated "In this age of excellent computing devices, the statistician should take a broader view and not select a narrow model prior to observing the sample items." Over the following years this estimator has been modified and the more recent version of this adaptive estimator, as described by Hogg and Lenth (1984), has excellent properties.

1.2.2 Rank Tests

After the paper by Hogg was published in 1967, the idea of using a selection statistic to modify a statistical method was developed further in a paper by Randles and Hogg (1973), which included an adaptive one-sample test and an adaptive two-sample test. In these tests the rank scores were selected based on selection statistics. Instead of using the sample kurtosis as a selection statistic, they used a measure of tailweight. Although these tests were adaptive in nature, they were not as powerful as traditional tests.

However, just two years later Hogg, Fisher, and Randles (1975) published an improved two-sample adaptive test. This test, which will be described in the next section, was the first practical and effective two-sample adaptive test. Although the test attracted considerable attention from statisticians, it appears to be rarely used by researchers. One problem with their adaptive test is that, because it is a rank-based test, it is not easy to generalize this approach to tests of significance of regression coefficients in more complex models.

In a series of articles published more than 20 years after the paper by Hogg, Fisher, and Randles (1975), several researchers used the same selection statistics to construct tests for a variety of situations. Büning (1996) proposed an adaptive test of equality of medians using data from a one-way layout. This test was based on an extension of Hogg's method of using selection statistics to select a set of rank scores. Two years later, Büning and Kossler (1998) proposed an adaptive test for umbrella alternatives and, in the following year, Büning (1999) proposed a test for ordered alternatives. Further extensions of the adaptive approach were made by Büning and Thadewald (2000), who proposed a location-scale test and by Büning (2002), who proposed

a test that could be used to test the null hypothesis that the distributions are equal against the general alternative that the distributions are not equal.

The tests proposed by Hogg and by Büning used selection statistics to determine the set of rank scores for the two-sample test. One small problem with this approach is that, if the selection statistics fall near the edge of a region corresponding to a set of rank scores, any small change in the data could change the selection statistics slightly, and this could result in the selection of an entirely different set of rank scores. This is undesirable because a small change in a single data value could result in a large change in the p-value. To remedy this situation, Ruberg (1986) proposed a continuously adaptive two-sample test and O'Gorman (1997) proposed a continuously adaptive test for the one-way layout. Using a different approach, Hall and Padmanabhan (1997) proposed several adaptive tests for the two-sample scale problem. They used a bootstrap testing approach with adaptively trimmed sample variances.

We should note that in the last 40 years there has also been work in the area of adaptive estimation. Yuh and Hogg (1988) proposed two adaptive regression estimators that rely on selection statistics to choose one of several robust regression estimators. Further work in the area of adaptive estimation was published by Hill and Padmanabhan (1991), who described the performance of some adaptive estimators when they were used with real data.

Although some material on adaptive confidence intervals and estimates is included in the last chapter of this book, we are primarily concerned with adaptive tests of significance. Further, we will focus on methods that can be used in a variety of testing situations. One such adaptive test utilizes a weighted least squares approach that we will now describe.

1.2.3 The Weighted Least Squares Approach

Before the year 2000, all of the adaptive tests were rank-based so they were limited to one-sample tests, two-sample tests, and tests for the one-way layout. A different approach was taken in 2001 when O'Gorman (2001) proposed a test that used an adaptive weighting method to assign weights to the observations so that the weighted observations could be used to test a subset of coefficients in a linear model. An improved version of this approach was proposed by O'Gorman (2002), and the book by O'Gorman (2004) described various applications of this method. With this approach the p-value was computed by using a permutation method. A few years later an adaptive multivariate test was proposed by O'Gorman (2006a), and this multivariate adaptive test was used in the analysis of repeated measures data by O'Gorman (2008a, 2010).

The adaptive tests proposed by O'Gorman prior to 2006 used a permutation method that required permutations of independent variables. An adaptive test that used permutation of residuals was proposed by O'Gorman (2006b), and this method was shown to be as effective as the test based on the permutation of independent variables. The advantages and disadvantages of various permutation methods will be addressed in Chapter 4. An improved adaptive weighting method that could be

used with univariate and multivariate data will be described in Chapters 2 and 3, and it will be used for most of the tests of significance.

1.2.4 Recent Rank-Based Tests

While most of the adaptive testing literature prior to 2000 focused on two-sample tests, some recent research has been published on one-sample adaptive tests. Lemmer (1993) suggested two adaptive tests for the median. Freidlin, Miao, and Gastwirth (2003) proposed an interesting and effective adaptive test for paired data. These authors use the p-value from a test of normality, rather than a measure of skewness or tailweight, as the selection statistic. They showed that their test is reasonably effective for moderate sample sizes. Baklizi (2005) proposed a continuously adaptive test for the median when the symmetry of the distribution is in doubt. It maintains its size and is relatively powerful. Most recently, Miao and Gastwirth (2009) proposed a test that uses the same score functions that were used by Freidlin, Miao, and Gastwirth (2003), but the test uses a measure of tail-heaviness as the selection statistic. This test will be described and evaluated in Chapter 11.

A different approach to robustifying and improving two-sample tests was taken by Neuhäuser, Büning, and Hothorn (2004). To construct their test, they used four sets of rank scores to produce four standardized linear rank statistics. Next, they computed the maximum of those four statistics as the overall test statistic, which is then used with a permutation method to compute the p-value. This test maintains its level of significance and has higher power than many of the traditional parametric and nonparametric tests. In addition, it has the advantage of not using any selection statistic. While it is not always classified as an adaptive test, it does achieve the same objective as the adaptive test. In Chapter 11 we will give a detailed description of this promising test.

1.3 THE ADAPTIVE TEST OF HOGG, FISHER, AND RANDLES

We will now consider the two-sample adaptive test that was proposed by Hogg, Fisher, and Randles (1975), which will be called the HFR test. Much of the early work in adaptive testing was based on the selection of an appropriate score function, which was a function of a selection statistic. Although Hogg (1967) used the sample kurtosis as a selection statistic, robust measures of asymmetry and tailweight were used in the HFR test. Their robust measure of asymmetry is calculated by combining the observations over both samples, sorting the observations, and then computing

$$Q_3 = \frac{\overline{U}_{.05} - \overline{M}_{.5}}{\overline{M}_{.5} - \overline{L}_{.05}},$$

where $\overline{U}_{.05}$ is the average of the upper 5%, $\overline{M}_{.5}$ is the average of the middle 50%, and $\overline{L}_{.05}$ is the average of the lower 5% of the observations. It should be noted that while Q_3 may be more robust than some other measures, $\overline{U}_{.05}$ and $\overline{L}_{.05}$ may be sensitive

to outliers. The robust measure of tail length is given by

$$Q_4 = \frac{\overline{U}_{.05} - \overline{L}_{.05}}{\overline{U}_{.5} - \overline{L}_{.5}},$$

where $\overline{U}_{.5}$ and $\overline{L}_{.5}$ are the averages of the upper and lower 50% of the combined samples, respectively. These two selection statistics are used to determine the most appropriate set of rank scores, as shown in the selection regions in Figure 1.1.

The general procedures for the HFR test are:

- Compute the selection statistics Q_3 and Q_4 using the combined observations.
- Use Q_3 and Q_4 to select the most effective rank scores based on the selection regions that are shown in Figure 1.1.
- Compute the test statistic using the selected scores and then find the p-value for the test based on those scores.

Hogg, Fisher, and Randles (1975) demonstrated that this test would maintain its significance level. They also showed, with a simulation study, that this adaptive test was often more powerful than traditional parametric and non-parametric tests.

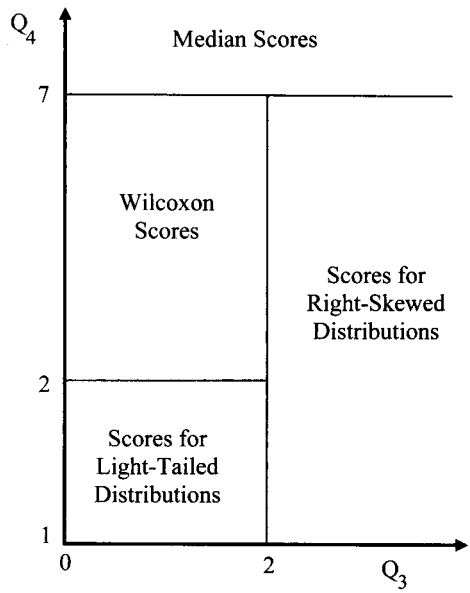


Figure 1.1 Selection regions for the HFR test.

To illustrate the selection method, suppose we obtained a data set for a two-sample test, with each sample having 20 observations, and we calculated the measure

of asymmetry to be $Q_3 = 1.1$, which indicated that the distribution was nearly symmetric. Next, we calculated the measure of tailweight as $Q_4 = 7.8$, which indicated that the distribution was long-tailed. With this set of selection statistics, we would have used the selection regions in Figure 1.1 to select the median scores. If the data were slightly skewed and we obtained $Q_3 = 1.5$ and $Q_4 = 3.5$ as selection statistics, the Wilcoxon scores would have been selected.

1.3.1 Level of Significance of the HFR Test

Hogg, Fisher, and Randles (1975) carefully constructed the adaptive HFR test so that it maintains its level of significance. To demonstrate this important property, let E_j be the event that the vector of selection statistics falls in region j and let R be the event that the null hypothesis is rejected. Then, for the jth region $P(R|E_j) \leq \alpha$ because these rank tests are distribution-free. Consequently, the HFR test will maintain its size because

$$P(R) = \sum P(R|E_j)P(E_j) \leq \sum \alpha P(E_j) = \alpha,$$

where the summations are over the four regions. Hence, even though we use the data to determine the rank scores, we find that the test has a significance level less than or equal to α. In addition, Hogg, Fisher, and Randles (1975) showed, using a simulation study with 15 observations per group, that the empirical significance level closely approximates α.

1.3.2 Comparison of Power of the HFR Test to the t Test

The real advantage of adaptive tests is that they often have greater power than the traditional test for many non-normal error distributions. To demonstrate this, we will compare the power of the HFR test to that of the pooled two-sample t test for several error distributions. The error distributions are listed in Table 1.1. We will compare the tests by means of a simulation study that uses 100,000 data sets for each distribution. This simulation program was written in FORTRAN to estimate the power of these tests. By writing the simulation program in FORTRAN we are able, in a reasonable amount of time, to analyze 100,000 data sets to obtain accurate estimates of the power.

For each data set we use 15 observations in each sample. For the first sample the observations are generated from one of the error distributions. For the second sample we add a constant δ to a random variable generated from the same distribution. After each data set is generated we perform the pooled two-sample t test and the HFR test. We count the number of times that the null hypothesis is rejected and calculate the proportion of rejections, which is the empirical power of the test.

These distributions will be described in much greater detail in Chapter 2. We note that the first four distributions are symmetric and that the other five are skewed. Some of these distributions have short tails while most of the others are long-tailed, and two are bimodal.

Table 1.1 Error distributions used in the simulation studies

Distribution	Description
Uniform	Uniform[0,1]
Normal	Standard normal $N(0,1)$
t_4	t distribution with d.f. = 4
Bimodal symmetric	Mixture of normals
RST $\alpha_3 = 1, \alpha_4 = 4.2$	Moderate skewness, low kurtosis
RST $\alpha_3 = 1, \alpha_4 = 8.4$	Moderate skewness, high kurtosis
RST $\alpha_3 = 2, \alpha_4 = 11.4$	High skewness, low kurtosis
RST $\alpha_3 = 2, \alpha_4 = 15.6$	High skewness, high kurtosis
Bimodal skewed	Mixture of normals

In order to make the power comparisons meaningful, we have set the shift (δ) of the random variables in the second population relative to those in the first population so that the powers of the t test would be near 80% for the nine error distributions. To meet this objective, we used $\delta = 1.0$ in a simulation study, with 15 observation in each sample for each of the 100,000 data sets that were generated. We obtained powers, which are displayed in Figure 1.2, for the t test and the HFR test with the nine error distributions. We note that the HFR test has greater power than the t test for most of the distributions but it has lower power for three of the distributions.

It is encouraging that the HFR test maintains its significance level and has greater power than the t test for several distributions, but it is discouraging to see a moderate power loss with the uniform, the normal, and the bimodal error distributions.

1.4 LIMITATIONS OF RANK-BASED TESTS

One problem with the Wilcoxon test, the adaptive HFR test, and other rank-based tests is that the testing methods make sense only when the data can be ranked. Consequently, rank-based tests are usually used to test equality of medians with two populations or to test equality of medians in a one-way layout. For example, Büning (1996) proposed an adaptive test for the one-way layout that used the HFR method of selecting rank scores. However, rank tests are difficult or impossible to perform with more complex models. For example, if we want to compare two groups and need to include one covariate, it may be difficult to find a suitable rank test.

To overcome this problem, O'Gorman (2001) proposed an adaptive test that did not use ranks. Instead, it uses an adaptive weighting scheme to improve the power of the tests. More recently, several variants of that method have been proposed to increase the power of the test and to allow it to be used with a wider variety of models. In this book we describe a new adaptive test that is powerful and flexible. The next section gives a brief description of this approach for the two-sample test. Most of

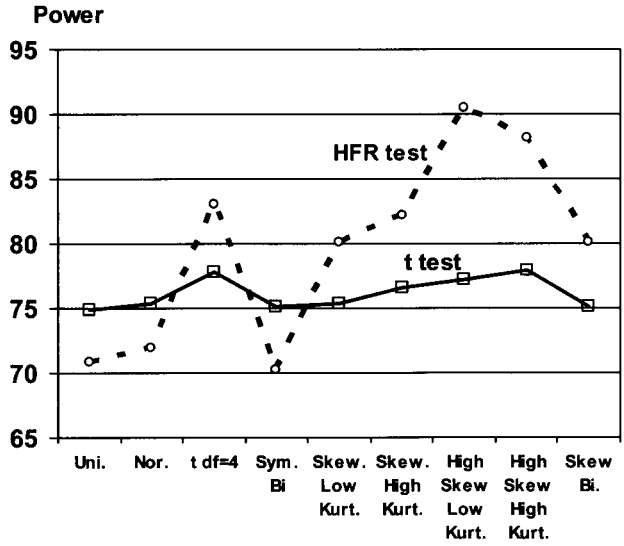


Figure 1.2 Power of the two-sample t test and the HFR test when 15 observations are generated for each sample. The t test is indicated by hollow squares and the HFR test is indicated by hollow circles.

the rest of this book is devoted to generalizing this testing procedure so that this new adaptive test can be used for most of the common testing situations.

1.5 THE ADAPTIVE WEIGHTED LEAST SQUARES APPROACH

The adaptive test that will be developed in this book is performed in two steps. In the first step the observations are weighed in a manner that produces residuals, in the weighted model, that are roughly normally distributed. In the second step a permutation method is used to compute a p-value. The details of the adaptive test will be described in subsequent chapters.

It may seem strange to use the weighted least squares method to normalize errors in regression models. In many books on regression analysis, the weighted least squares method is used to ensure that the errors have the same variability. See Rawlings, Pantula, and Dickey (1998, Chapter 12) for a description of the weighted least squares method in regression. In this book we are primarily interested in adaptive methods for non-normal data, so we are interested in weighting the observations to normalize the errors.

An example may illustrate the basic approach used in the adaptive weighted least squares test. In an experiment that was described by Koziol *et al.* (1981), mice were injected with colon carcinoma cells in order to determine the effectiveness of several immunotherapy regimens. Five days later the mice were randomly assigned to Group

1 or to Group 2. Mice in Group 1 received injections of tissue culture medium around the tumor while mice in Group 2 received injections of normal spleen cells, immune RNA, and tumor antigen. The tumor volumes (mm^3) at day 13 are given in Table 1.2 for the 10 mice in these two groups. In Koziol *et al.* (1981), Group 1 and Group 2 are labeled as Group A and Group C, respectively.

Table 1.2 Tumor volumes, in mm^3, for 2 groups of mice measured on day 13 of the experiment and the adaptive weights. Group 1 is the group that received the tissue culture medium and Group 2 is the group that received the spleen cells, immune RNA, and tumor antigen.

Group 1			Group 2		
Obs	Volume	Weight	Obs	Volume	Weight
1	217.6	1.118	11	186.2	1.107
2	176.6	1.044	12	196.6	1.143
3	196.1	1.142	13	191.3	1.129
4	205.9	1.147	14	129.6	0.625
5	196.0	1.142	15	420.0	0.400
6	225.1	1.081	16	32.0	0.511
7	274.7	0.800	17	55.0	0.466
8	202.5	1.149	18	84.7	0.470
9	205.8	1.147	19	258.8	0.867
10	225.0	1.082	20	176.4	1.043

To show the relative effectiveness of these regimens, the volumes are displayed in Figure 1.3. This dot plot shows that there are several outliers present in Group 2 and that the tumor volumes in Group 1 tend to be larger than those in Group 2. However, if we perform a two-sample test, the pooled t test gives a p-value of $p = 0.292$ and the unequal variance t test gives a p-value of $p = 0.304$. These large p-values are due, in part, to the large variability that is present in Group 2.

In the adaptive WLS test we give reduced weights to the observations that are extreme. The adaptive weights are given in Table 1.2 for the 20 mice. Note that mouse 15 was given a weight of 0.400 because the tumor volume was so large, and mice 16, 17, and 18 were given low weights because the tumor volumes were quite small. In this way the adaptive test reduces the influence of the outliers. We then used a permutation method to find a p-value of $p = 0.061$, which is much smaller than the p-values found with the pooled and unequal variance t tests.

However, it is important not to place too much importance on the difference in p-values for this one data set because we do not know if the distributions are different. The weighting procedures and permutation methods will be described in great detail in Chapters 2 and 3. In the remainder of this chapter we will briefly describe the significance level and the power of this adaptive test.