Modern Robust Statistical Methods: Basics with Illustrations Using Psychobiological Data

Rand R. Wilcox¹*, Douglas A. Granger², Florence Clark³

¹Dept. of Psychology, University of Southern California, Los Angeles, CA 90089-1061, United States
²Center for Interdisciplinary Salivary Bioscience Research, Johns Hopkins University, School of Nursing, Bloomberg School of Public Health, and School of Medicine
³Division of Occupational Science & Occupational Therapy, University of Southern California

*Corresponding Author: rwilcox@usc.edu

Copyright ©2013 Horizon Research Publishing All rights reserved.

Abstract Psychological studies in general, and psychobiological studies in particular, routinely use a collection of classic statistical techniques aimed at comparing groups or studying associations. A fundamental issue is whether violating the basic assumptions underlying these methods, namely normality and homoscedasticity, can result in relatively poor power or miss important features of the data that have practical significance. In the statistics literature, hundreds of papers make it clear that under general conditions the answer is yes and that routinely used strategies for dealing with violations of assumptions can be unsatisfactory. Moreover, a vast array of new and improved techniques is now available for dealing with violations of assumptions, including more flexible methods for dealing with curvature. The paper reviews the major insights regarding standard methods, explains why some seemingly reasonable methods for dealing with violations of assumptions are technically unsound, and then outlines methods that are technically correct. It then illustrates the practical importance of modern methods using data from the Well Elderly II study.

Keywords Psychobiological studies, robust statistical methods, outliers, curvature, heteroscedasticity, smoothers, biomarkers, depressive symptoms, meaningful activity

1 Introduction

Classic statistical techniques, routinely used in psychological studies, are based in part on two basic assumptions: normality and homoscedasticity (equal variances). A well-known view is that these techniques are robust when either of these assumptions is violated. But it is well known in the statistics literature that, under general conditions, commonly used methods can be highly unsatisfactory (e.g., [18, 21, 22, 36, 38, 42, 43]). It is evident, however, that psychobiological studies do not take the advances and insights into account.

Briefly, standard methods are robust in terms of Type I errors when comparing groups that have identical distributions, or when dealing with regression and the variables under study are independent. So a positive feature of classic methods based on means is that if they yield a significant result, it is reasonable to conclude that distributions differ in some manner. And if when using Pearson’s correlation, if Student’s t rejects, it is reasonable to conclude that there is some type of association. Of course, if distributions differ or there is an association, standard methods might continue to perform well in terms of power, accurate confidence intervals, as well as providing a reasonable characterization of how groups differ or the association among variables. But to assume that these methods perform in a relatively effective manner is not supported by hundreds of published papers.

Transforming data is a well-known strategy for dealing with violations of assumptions, but by modern standards this approach is relatively ineffective for reasons that are reviewed and illustrated.

There have been three major insights that have serious implications about the accuracy of methods based on means in particular and least squares regression more generally. Simultaneously, many new and improved techniques have been derived that deal with known concerns. The bulk of these methods center around four major advances: the theory of robustness (e.g., [18, 22, 38]) improved methods for dealing with skewed distributions and outliers (e.g., [21, 25, 36, 42, 43]), hypothesis testing methods that allow heteroscedasticity [42, 43] and the free software R [31]. Consequently, the goal in this paper is to review the basics of modern robust methods and to illustrate, using data from the Well Elderly II study [23] that they can make a practical difference in psychological studies. A point worth stressing is that modern methods do more than provide relatively high power and accurate confidence intervals: they provide new and interesting perspectives that help foster a deeper understanding of data.
2 Summary of Three Major Insights

As previously indicated, there have been three major insights regarding classic methods based on means and the least squares regression estimator. Briefly, they deal with 1) the impact of heavy-tailed distributions and outliers, 2) skewed distributions and the central limit theorem, and 3) heteroscedasticity. The immediate goal is to describe and illustrate each of these issues.

2.0.1 Heavy-tailed Distributions and Outliers

A small departure from a normal distribution can have a large impact on the population variance, which in turn could mean relatively poor power when comparing means. A classic illustration stems from the mixed (or contaminated) normal distribution as discussed by Tukey (1960). The particular mixed normal distribution he considered is a situation where with probability 0.9 an observation is sampled from a standard normal, otherwise an observation is sampled from a normal distribution with mean 0 and standard deviation 10. Figure 1 shows the standard normal distribution, which of course has variance one, and the mixed normal. Despite the obvious similarity between the two distributions, the mixed normal has variance 10.9. The mixed normal is an example of a heavy-tailed distribution meaning that the tails are thicker than a normal distribution.

Figure 1. Shown are the standard normal and mixed normal distributions. Despite the obvious similarity between the two distributions, the variance of the mixed normal is 10.9

Figure 2 illustrates the impact of a heavy-tailed distribution on power. In the left panel are two normal distributions, each having variance one with means equal to 0 and 1. The power of Student’s t, when testing at the .05 level and with sample sizes of 25 per group, is .96. But in the right panel, which shows two mixed normals, again having mean 0 and 1, power is only .28. Put another way, random samples from heavy-tailed distributions are characterized by outliers and even a single outlier has the potential of inflating the sample variance, which in turn can result in poor power. The presence of outliers does not necessarily spell disaster, but simply ignoring the potential impact of outliers cannot be recommended.

2.0.2 Skewed Distributions and the Central Limit Theorem

Next consider the central limit theorem. A common misconception is that with a sample size of 25 or more, normality can be assumed. So in particular, $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

is assumed to have a Student’s t distribution with $n - 1$ degrees of freedom, where $n$ is the sample size, $\bar{X}$ is the usual sample mean and $\mu$ is the population mean. This view stems from studies where sampling is from light-tailed distributions, roughly meaning that outliers are relatively rare. However, these early results missed two fundamental points. The first has to do with situations where a skewed distribution tends to have outliers. Now, more than 100 observations might be needed so that the sample mean has, approximately, a normal distribution. Second, and perhaps more importantly, is the implicit assumption that if the sample mean has a normal distribution, Student’s t will be reasonably accurate in terms of Type I error probabilities and confidence intervals. Even with a skewed, light-tailed distribution, where the sample mean has approximately a normal distribution, a sample size of over 200 might be needed to get accurate results when using Student’s t (e.g., [41, 42, 43]). For a skewed, heavy-tailed distribution (outliers are common), a sample size of 300 or more might be needed (e.g., [42, 43]).

Data from the Well Elderly II study ([23] are used to illustrate the impact of skewed distributions on the distribution of $T$. The participants were men and women aged 60 to 95 years (mean age 74.9). The sample size is $n = 254$ after missing observations were eliminated. Their study was generally aimed at testing the hypotheses that a six-month lifestyle, activity-based intervention leads to reduced decline in physical health, mental well-being and cognitive functioning among ethnically diverse
older people. One of the variables was salivary cortisol upon awakening, 30-60 minutes later, 5 hours later, and 5 hours later. Cortisol has been found to be associated with various measures of psychological stress and is generally considered to be a valuable marker of the hypothalamus-pituitary-adrenal (HPA) axis (e.g., [15]). Compliance with the sampling schedule was monitored by recording the times when the sample was taken. Samples were assayed for cortisol using a highly sensitive enzyme immunoassay without modifications to the manufacturers recommended protocol (Salimetrics, State College, LLC). The test uses 25ul test volume, ranges in sensitivity from .007 to 3.0 ug/dl, and has average intra- and inter-assay coefficients of variation of 4.13% and 8.89%, respectively.

Figure 3 shows a boxplot of the cortisol data. As is evident, at all four times, the data are skewed with outliers. Boxplots for DHEA (dehydroepiandrosterone), which is also considered to be a valuable marker of the HPA axis, and salivary alpha amylase, not shown here, are similar to those in Figure 3. (Salivary secretion of alpha amylase has been proposed as an indicator of plasma catecholamine modifications under a variety of conditions. Catecholamines play an important role in the body's physiological response to stress.)

The skewed distribution in Figure 4 is a bootstrap-t approximation of the distribution of $T$ based on the cortisol data upon awakening. Based on a sample of $n$ observations, the bootstrap-t method randomly samples with replacement $n$ observations from the observed data, computes $T$ and then repeats this $B$ times. Here, $B = 2000$ was used. (In effect, perform a simulation study on the observed data.) Also shown is the distribution of $T$ under normality. As can be seen the left tails differ substantially despite having a sample size of $n = 254$. Under normality, when testing a two-sided hypothesis at the .05 level, the lower .025 critical value would be $-1.97$. But the approximation indicates that it should be $-3.03$. Put another way, if we assume normality and reject when $T$ is less than $-1.97$, the probability of a Type I error, when the null hypothesis is true, is .025. But the approximation indicates that the actual Type I error probability is .071, nearly three times larger than intended. In the right tail we see the opposite problem. The probability of rejecting is estimated to be .008, substantially less than the intended level of .025, which has implications about low power when the null hypothesis is false and the actual value of $\mu$ is greater than the hypothesized value.

A valid criticism of this illustration is that the bootstrap-t approximation of the distribution of $T$ might be inaccurate. Results summarized in [43] indicate that indeed there are practical concerns: problems with Student's t are probably worse than indicated. In particular, in all likelihood, the left tail does not extend out far enough and the actual Type I error is most likely higher than indicated. The bootstrap-t method can improve control over the Type I error probability, but in general practical problems remain.

In the Well Elderly II study, measures were taken again six months after intervention. Boxplots of cortisol, DHEA and alpha amylase are similar to those in Figure 3. For cortisol measured upon awakening, the Type I error probability was estimated to be .086 for the lower tail when testing at the .025 level, using Student's t, and for the upper tail it was estimated to be .004. So for a two-sided hypothesis, the estimate is that when testing at the .05 level, the actual level is .090 and it is probably higher based on known results regarding the bootstrap-t approximation that was used.

A positive feature of Student's t test is that for symmetric heavy-tailed distributions, the actual Type I error probability tends to be less than or equal to the nominal level. But it can be substantially less than the nominal level, which exacerbates power problems. For symmetric, light-tailed distributions, including normal distributions as a special case, control over the Type I error probability is generally good. This helps explain why, when comparing two groups that have distributions with the same amount of skewness, past simulation studies found that the Type I error probability is controlled reasonably well. (If two random variables have the same skewness, their difference will have a symmetric distribution.) But more recent studies, where distributions differ in skew-
ness, illustrate that now control over the Type I error probability can be unsatisfactory. In practical terms, if the goal is to test the hypothesis of identical distributions, Student’s t is satisfactory in terms of controlling the probability of a Type I error. It might provide accurate information about the population mean, but under general conditions this is not the case.

2.0.3 Heteroscedasticity

When using a hypothesis testing method that assumes homoscedasticity, heteroscedasticity further exacerbates the problems already described and illustrated. One basic reason is that under general conditions, if there is heteroscedasticity, homoscedastic methods are using the wrong standard error. Cressie and Whitford (1986) describe general conditions where, when comparing two independent groups, Student’s t is not even asymptotically correct. That is, when testing at the .05 level, for example, a basic requirement of any method is that the actual Type I error probability should converge to .05 as the sample size gets large. But under general conditions, Student’s t does not enjoy this property. Moreover, the more complicated the design, the more serious are problems associated with heteroscedasticity when using a homoscedastic method (e.g., Wilcox, 2012a). For example, under normality, when comparing two independent groups, Student’s t controls the probability of a Type I error reasonably well when equal sample sizes are used. But for the ANOVA F test, when comparing four groups, this is no longer the case. Again, if the goal is to test the hypothesis that groups have identical distributions, Type I errors are controlled reasonably well via the ANOVA F test. But if the goal is to control Type I errors when testing the hypothesis of equal means, the ANOVA F test can be highly unsatisfactory.

3 Some Unsatisfactory Strategies

3.0.4 Detecting Outliers

When checking for outliers, methods based on the mean and standard deviation are known to be highly unsatisfactory because this approach results in masking, roughly meaning that the very presence of outliers can cause them to be missed (e.g., [36, 42]). Briefly, outliers inflate the sample variance so that even extreme outliers can be missed. For univariate data there are two basic outlier detection techniques that avoid masking. The first is a boxplot and the other is the MAD-median rule. To describe the MAD-median rule, let $X_1, \ldots, X_n$ be $n$ observations and let $M$ be the usual sample median. The median absolute deviation (MAD) is equal to the median of $|X_1 - M|, \ldots, |X_n - M|$. It represents a measure of variation that is insensitive to outliers, a key feature needed to avoid masking. The MAD-median rule declares the value $X$ an outlier if

$$\frac{|X - M|}{MAD} > 2.24.$$  

(Under normality, MAD/.6745 estimates the usual population standard deviation.) Compared to the boxplot, the MAD-median rule is better at avoiding masking. Nevertheless, there are practical reasons for still using a boxplot, but the details go beyond the scope of this review. For multivariate data, Mahalanobis distance is unsatisfactory as well: it suffers from masking. For a recent summary of more satisfactory methods for detecting multivariate outliers, see Wilcox (2012b).

3.0.5 Dealing with Heteroscedasticity

A seemingly natural strategy for dealing with heteroscedasticity is to test the hypothesis that there is homoscedasticity and use a homoscedastic method if a non-significant result is obtained. However, five published papers (summarized in Wilcox, 2012a) have studied this strategy and all five came to the same conclusion: this approach is unsatisfactory. The simple explanation is that the power of the methods used to test the homoscedasticity assumption is not high enough to detect situations where the assumption should be abandoned. Currently, a better strategy is to always use a method that allows heteroscedasticity. Little is lost if in fact there is homoscedasticity, but much can be gained in terms of accurate probability coverage and power. This remains the case when dealing with regression [26].

3.0.6 Transforming Data

There are two fundamental limitations regarding simple transformations (e.g., [8, 33]). First, transforming data does not deal effectively with outliers. In some situations the number of outliers is decreased, but in other situations the number stays about the same and can even increase. The fact that even a single outlier might result in poor power renders simple transformations to be an effective method. Second, typically, but not always, distributions remain substantially skewed.

Consider again the cortisol data in Figure 3. Taking logs, the number of outliers drops from 19 to 13 and a boxplot again indicates a skewed distribution. So a positive feature of taking logs is a reduction in the number of outliers. But a serious concern is that outliers remain that can negatively impact power. And it is unclear that problems associated with skewed distributions have been adequately addressed.

3.0.7 Discarding Outliers and Applying Methods for Means Using the Remaining Data

Based on standard training, a seemingly reasonable way of dealing with outliers is to discard them and apply a method for means using the remaining data. But even if a heteroscedastic method is used, this strategy is technically unsound when dealing with any dependent variable and can yield highly inaccurate results even with a large sample size. The reason is that the derivation of the standard error is no longer valid. More importantly, using a correct estimate of the standard can make a substantial difference when analyzing data (e.g., [42]. A crude explanation is that when extreme but valid values are removed, the remaining observations are no longer independent, which results in having to use alternative methods for estimating standard errors. Technically sound methods are available (e.g., [21, 38, 21, 43]),
but they are not remotely obvious based on standard training.

4 Technically Sound Methods for Dealing with Skewed Distributions and Outliers

There are several ways of dealing with skewed distributions and outliers in a technically sound manner with each being sensitive to different features of the data and providing different perspectives on how groups compare and how variables are related.

4.0.8 The Median and Trimmed Means

One possibility is to compare groups using the usual sample median. It is highly insensitive to outliers because after putting observations in ascending order, all but one or two values are trimmed. Also, if the goal is to characterize the typical response, the median is arguably a better choice than the mean when dealing with a skewed distribution. However, a concern is that because the usual sample median trims nearly all of the observations, power might be relatively low. This is the case under normality. But even two centuries ago, Laplace was aware of conditions where the median has a smaller standard error than the mean (Hand, 1990). In modern terms, when dealing with a sufficiently heavy-tailed distribution where the number of outliers tends to be large, comparing medians can result in more power than methods based on the mean.

A strategy for dealing with the relatively large standard error associated with the median is to simply trim less. But how much trimming should be done? One approach is to trim by an amount that guards against a reasonably large number of outliers yet competes well with the mean under normality. Based on this view, Rosenberger and Gasko (1983) concluded that a 20% trimmed mean is a relatively good choice. That is, the lower 20% and the upper 20% of the values are trimmed and the remaining observations are averaged. Like the median, a trimmed mean can provide a better reflection of the typical response when dealing with a skewed distribution. Moreover, when testing hypotheses, both theory and simulations indicate that using a trimmed mean can substantially reduce the problems associated with the mean in terms of Type I error probabilities that are associated with skewed distributions.

As an illustration, consider again cortisol measured upon awakening. With a 20% trimmed mean, the probability of a Type I error associated with the lower tail, when testing at the .025 level, is estimated to be .028 and for the upper tail the estimate is .018, again using the bootstrap-t method. So for a two-sided test, when testing at the .05 level, the actual Type I error probability is estimated to be .046. Moreover, when working with a 20% trimmed mean rather than the mean, the bootstrap-t method provides a substantially better method for assessing the actual Type I error probability.

A theoretically correct method for estimating the standard error of a trimmed mean was first derived by [40]. It involves, in part, a Winsorized variance. To explain, suppose that when computing the trimmed mean, the two lowest observations are trimmed. Then Winsorizing means that rather than trim these observations, they are set equal to the lowest value not trimmed. Similarly, if when computing the trimmed mean the two largest values are trimmed, Winsorizing means the two largest value are set equal to the largest value not trimmed. The Winsorized variance is the variance of the Winsorized values. More formally, for a 20% trimmed mean, if $s^2_\omega$ represents the 20% Winsorized variance, the estimate of the squared standard error of the 20% trimmed mean is $s^2_\omega/(.6n)$, where $n$ is the sample size before any observations are trimmed.

Comparing 20% trimmed means can result in much higher power when dealing with heavy-tailed distributions. Consider again the right panel of Figure 2 where Student’s t has power .28. Using instead the method derived by [48], power is .85.

Today, all of the usual ANOVA designs can be analyzed using trimmed means (Wilcoxa, Wilcoxb). These methods have been studied extensively and found to have practical advantages over methods based on means. In particular, they perform well when there is heteroscedasticity. Note, however, that for skewed distributions, comparing means is not the same as comparing trimmed means or medians. So despite having smaller standard errors, it is possible for a method based on means to reject when methods based on trimmed means or medians do not. Also, by design, methods based on trimmed means are primarily sensitive to only differences among the population trimmed means. In contrast, methods based on means are sensitive to differences in skewness, and homoscedastic methods are sensitive to heteroscedasticity, which again might mean more power than methods based on a trimmed mean.

It is stressed that although the median belongs to the family of trimmed means, inferences based on the sample median require special techniques, particularly when there are tied (duplicated) values. Tied values wreak havoc on methods aimed at estimating the standard error of the sample median. When testing hypotheses, a very effective method for dealing with tied values is a percentile bootstrap method (e.g., [43]).

An alternative to trimmed means is to empirically check for any outliers and down weight or eliminate them and then average the remaining values. An important special class of estimators that effectively uses this strategy is the class of robust M-estimators. Expressions for the standard error of an M-estimator have been derived, but when testing hypotheses, the resulting test statistics tend to perform poorly when dealing with skewed distributions. However, percentile bootstrap methods can be used to test hypotheses, which control Type I error probabilities relatively well even when there is heteroscedasticity. Details are summarized in [43].

4.0.9 Rank-Based Methods

Another way of dealing with outliers is to use a rank-based or nonparametric method, with the understanding that in general, they are designed to be sensitive to different features of the data compared to methods
Based on measures of location. A point that should be stressed is that all of the classic, routinely taught rank-based methods have been improved substantially.

To elaborate, consider the Wilcoxon–Mann–Whitney (WMW) test. The WMW test is sometimes suggested for comparing medians, but it is inappropriate for this purpose under general conditions (e.g., [14]). That is, if \( X \) and \( Y \) are two independent variables, the WMW test is not designed to test \( H_0: \theta_x = \theta_y \), where \( \theta_x \) is the population median associated with \( X \). Rather, it is based on a direct estimate of \( p = P(X < Y) \), the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second. If the groups do not differ, then in particular \( p = 0.5 \). Certainly information about \( p \) is one way of characterizing how groups differ in a simple and useful manner. When the two groups have identical distributions, the WMW test uses a correct expression for the standard error. But when the distributions differ, this is no longer the case. So in essence, it provides a reasonable test of the hypothesis that distributions are identical, but it is not a satisfactory method for making inferences about \( p \). Methods that use a correct estimate of the standard error when distributions differ have been derived (e.g., [2, 5]). Similar problems are associated with the Kruskall–Wallis test and Friedman’s test, but modern techniques deal effectively with known concerns including issues related to tied values (e.g. [2, 43]).

5 Regression and Measures of Association

Least squares regression and Pearson’s correlation inherit all of the practical problems associated with methods aimed at comparing means and new problems are introduced. A vast array of new methods has been derived that can make a substantial difference in terms of both power and our understanding of any association that might exist.

5.0.10 Heteroscedasticity

In the context of regression, homoscedasticity means that the conditional variance of the dependent variable \( Y \), given some value for the independent variable \( X \), does not depend on the value of \( X \). Independence implies homoscedasticity, but when there is an association, there is, of course, no reason to assume that there is homoscedasticity. Indeed, the homoscedasticity assumption is imposed simply to avoid a technical issue: estimating the standard errors of the least squares estimates of the parameters when there is heteroscedasticity. In recent years, several methods for estimating standard errors, when there is heteroscedasticity, have been derived. No single estimator is always best, but the so-called HC4 estimator appears to perform relatively well (e.g., [7, 16]) and it can make a practical difference in terms of achieving accurate confidence intervals for the slopes and intercept. The HC4 estimator is readily applied using extant software, as indicated later in the paper.

5.0.11 Outliers

Outliers are a concern when using least squares regression or Pearson’s correlation because even a single outlier can destroy power or yield a highly misleading summary of the bulk of the data. The presence of outliers does not necessarily mean that ordinary least squares (OLS) will perform poorly, but ignoring the practical concerns associated with outliers can yield a highly misleading summary of the data.

Two general types of outliers play a role in regression. The first is leverage points, meaning outliers among the explanatory (independent) variable. The other is outliers among the dependent variable. One reason outliers among the dependent variable are a concern is that they inflate the standard error of the least squares estimator, which in turn can result in relatively poor power. A simple way of improving the OLS estimator is to consider the impact of removing leverage points, taking care to use a method that avoids masking. When there is more than one independent variable, no single outlier detection technique dominates, but one that performs relatively well is a projection method (e.g., [43], section 6.4.9). Removing leverage points does not invalidate the use of extant heteroscedastic methods; the derivation of the method for estimating standard errors remains valid.

But when removing outliers among the dependent variable, now the derivation of the estimates of the standard errors is no longer valid. There are, however, technically sound techniques for dealing with this problem. First, use a robust regression estimator that deals with outliers among the dependent variable in a reasonably effective manner. Many such estimators are now available ([43], Chapter 10). Two that perform relatively well are the Theil [39] and Sen [37] estimator and the MM-estimator derived by Yohai [47], but arguments for considering other estimators can be made. (For recent results on handling tied values among the dependent variable, see [44].) Next, test hypotheses using a method that allows heteroscedasticity. Typically a basic percentile bootstrap method performs well, which can be applied with extant software as described below.

5.0.12 Measuring the Strength of an Association

One way of dealing with outliers when measuring the strength of an association is to use some analog of Pearson’s correlation that down weights or eliminates outliers among the marginal distributions. Two classic examples are Spearman’s correlation and Kendall’s tau. Another is the Winsorized correlation. It Winsorizes the marginal distributions, after which Pearson’s correlation is computed.

Two points should be stressed. First, classic methods for testing the hypothesis of a zero correlation assume homoscedasticity. That is, when they reject, the main reason might have to do more with heteroscedasticity rather than the strength of the association. Methods for dealing with heteroscedasticity have been derived (e.g., [43]), but no details are given here. Second, although Kendall’s tau and Spearman’s correlation guard against outliers among the marginal distributions, properly placed outliers can have a substantial impact on
their values because they do not take into account the overall structure of the data when dealing with outliers. Roughly, a point can be highly unusual relative to the bulk of the points even when no outliers among the marginal distributions are detected. (See [42], section 14.1 for illustrations.) Methods for dealing with this problem are available with the so-called skipped correlation currently a relatively good choice (e.g., [27, 42, 43]).

Another general approach to measuring the strength of an association is to first fit a regression line and then use what is called explanatory power, which contains Pearson’s correlation as a special case. Let $\hat{Y}$ be the regression estimate of $Y$ given the value of some independent variable $X$. Let $\sigma^2(\hat{Y})$ be the variance of the predicted $Y$ values and let $\sigma^2(Y)$ be the variance of the observed $Y$ values. Then explanatory power is

$$\eta^2 = \frac{\sigma^2(Y)}{\sigma^2(\hat{Y})}$$

If $\hat{Y}$ is based on the OLS estimator, $\eta^2$ reduces to $\rho^2$, the usual coefficient of determination. A robust analog of $\eta^2$ is obtained by replacing the variance with some measure of variation that is less sensitive to outliers in conjunction with a robust regression estimator. A simple alternative to the usual variance is the Winsorized variance, but arguments for considering two other measures of variation (the biweight and percentage bend measures of variation) can be made [43]. A possible appeal of explanatory power is that it is readily applied when dealing with curvature in a non-parametric fashion as discussed next.

### 5.0.13 Curvature

Basic training suggests ways of dealing with curvature using well-known parametric models. But there have been many advances regarding how to deal with curvature in a more flexible manner (e.g., [10, 11, 13, 17, 20, 42, 43]). Moreover, experience with these more modern methods suggests that often they can make a substantial difference in our understanding of associations. And they can help discover associations that would otherwise would be missed, as illustrated below. The sobriquet for these techniques is smoother, many of which have been derived. An early and relatively effective method was derived by Cleveland [3] for the case of a single independent variable and was later generalized to more than one independent variable by Cleveland and Devlin (1988). To provide a crude indication of the basic strategy, momentarily assume the goal is to predict $Y$, given that $X = 6$, based on the observed pairs of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$. Cleveland’s method uses weighted least squares with the weights $w_1, \ldots, w_n$ determined by how close $X_i$ is to 6. The $X_i$ values relatively far from 6 get no weight. That is, they are ignored. Note that this process can be done for a range of $X$ values and in particular for each observed $X_i$. This yields $n$ pairs of points, say $(X_1, \hat{Y}_1), \ldots, (X_n, \hat{Y}_n)$. Plotting these points yields the simplest version of the smooth derived by Cleveland [3]. Included is a method for down weighting outliers among the dependent variable $Y$. There are important alternative strategies but the many details are too involved to give here. (For a generalization to multiple independent variables, see [4].)

### 5.0.14 Software

An important practical point is that modern robust methods can be applied with extant software. Easily the best software, in terms of gaining access to the many new techniques that have been derived, is the free software R, which can be downloaded from www.R-project.org. R is tremendously powerful and contains all of the usual methods that one would expect to find. For Illustrations on how to apply modern robust methods using R, that cover a relatively wide range of situations, see [42, 43].

### 6 More Illustrations

As a simple illustration of modern statistical techniques, we consider the association between awakening cortisol and DHEA in older adults, again using the Well Elderly data. As previously indicated, there are outliers associated with both variables, but to illustrate some technical issues, outliers are momentarily retained. Then Pearson’s correlation is $r = .136$ with $p = .001$ (based on Student’s $t$) indicating that there is an association. From basic training, this result suggests that as cortisol increases, so does the expected value of DHEA. If, however, one uses the HC4 method for dealing with heteroscedasticity (via the R function pcorhc4), $p = .49$. Moreover, a test of the hypothesis that there is homoscedasticity (using the R function qhomt) rejects ($p = .016$), the main point being that a homoscedastic method can give a substantially different result compared to one designed to deal with heteroscedasticity. (The reverse also happens where a heteroscedastic method is highly significant but a homoscedastic method is not.)

A fairly obvious way of dealing with outliers is to use Spearman’s correlation, which is estimated to be $r_s = .32$ for the situation at hand, with $p < .001$ based on the usual Student’s $t$ test, which assumes homoscedasticity. This result is fairly consistent with one reported in [1] where $r_s = .43$, $p = .003$. Switching to a heteroscedastic method (via the R function corhc), again Spearman’s correlation is significant ($p < .001$). Switching to a skipped correlation, with the goal of taking into account the overall structure of the data when dealing with outliers, the correlation is now .23 suggesting that outliers are having some impact on Spearman’s correlation, but again a significant result is obtained using a heteroscedastic method (using the R function scorci, $p < .001$).

So there are multiple indications that cortisol and DHEA are associated and that the association is positive. But one appeal of modern methods is that a more detailed understanding of the nature of the association is possible via a smoother. Figure 5 shows a smooth (created with the R function lplot) where cortisol is used to predict DHEA. Also shown is the usual least squares regression line. (There are three extreme outliers among the cortisol values that were eliminated.) Note that there appears to be a distinct bend close to .4 on the $x$-axis. A test of the hypothesis that the regression line
is straight rejects ($p < .001$). Also, splitting the data in terms of whether cortisol is greater than or less than .4, the resulting regression lines have significantly different slopes. Now a positive association is found for cortisol less than .4 (using the R function regci), but no association is found when cortisol is greater than .4. So the general conclusion is that there is a positive association when cortisol is relatively low, but for higher levels of cortisol there is no indication of an association. Notice that Spearman’s correlation paints a decidedly different picture. It underestimates somewhat the strength of the association when cortisol is low and misleadingly suggests that there is positive association when cortisol is high.

![Figure 5](image)

**Figure 5.** Estimated regression line for predicting DHEA given cortisol using least squares and a smoother

In some situations, simply removing leverage points and using OLS gives very similar results to more modern regression estimators. But caution is warranted because outliers among the dependent variable can result in relatively poor power when using OLS. That is, a robust estimator is needed. The next two illustrations demonstrate both points.

First consider the association between a measure of meaningful activities and the cortisol awakening response (CAR), which is the change in cortisol measured at the time of awakening and again 30-60 minutes later, again using the Well Elderly data. Pruessner et al. [30] were the first to propose that the repeated assessment of the salivary cortisol increase after awakening might represent a useful and easily measured index of cortisol regulation. Exhibiting an absence or an exacerbation of this increase is associated with several adverse psychological and physiological outcomes (e.g., [28, 29]). Meaningful activities were measured with the Meaningful Activity Participation Assessment (MAPA) instrument, which was studied [9] and found to be a reliable and valid measure of meaningful activity, incorporating both subjective and objective indicators of activity engagement. For the version of MAPA used here, higher scores reflect a greater activity participation, with an individual’s MAPA score consisting of the sum of 29, 7-point Likert scales. The range of observed scores was 0-97 with 14 outliers based on the MAD-median rule.

Applying OLS with the usual Student’s $t$ test, no association is found ($p = .88$), and a heteroscedastic method fails to reject as well ($p = .96$). But removing leverage points (using a MAD-median rule), now a significant result is obtain ($p = .005$). Using the Theil–Sen estimator, $p = .08$ when leverage points are retained and $p = .007$ when leverage points are removed. So even when using a modern regression estimator, it can be important to check the impact of removing leverage points.

To illustrate the potential impact of outliers among the dependent variable, consider the goal of predicting the CAR given a MAPA score. Using a smoother in a manner that deals with outliers yields the solid line shown in Figure 6. The nearly identical dashed line is the Theil–Sen regression line. The nearly horizontal (dotted) line is the least squares regression line. Using the usual Student’s $t$ test in conjunction with the least squares regression line yields a nonsignificant result ($p = .80$) and removing leverage points again yields a nonsignificant result ($p = .94$). But using the Theil–Sen estimator, $p = .033$, and when leverage points are removed, $p = .023$.

![Figure 6](image)

**Figure 6.** Regression lines for predicting MAPA given CAR. The nearly horizontal line is the least squares regression line.

Finally, consider MAPA as the dependent variable, with CAR and a measure of depressive symptoms as the independent variables. (Now, another form of MAPA is used where higher scores reflect a greater activity satisfaction rather than participation.) Here, depressive symptoms were measured with the Center for Epidemiologic Studies Depression Scale (CESD). The CESD [32] is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Foley et al., 2002; Lewinsohn et al., 1988). Applying OLS in the usual way, the slope associated with CESD is significant ($p < .001$), but the slope associated with the CAR is not ($p = .69$). Dealing with heteroscedasticity via the HC4 estimator and removing leverage points, again the slope associated with the CAR is not significant ($p = .79$). Using the MM-estimator to deal with non-normality, again removing leverage points, once more a non-significant result is obtained for the CAR ($p = .68$). So the results find no association between the CAR and activity satisfaction when taking
depressive symptoms into account.

However, look at Figure 7 which shows a smooth for predicting MAPA given CESD and the CAR. Note that there appears to be a distinct bend where the CAR is approximately 18. That is, there appears to be curvature with the nature of the association changing when depressive symptoms are relatively high. Using only the data where CAR is less than 18, now the slope associated with CAR is significant ($p = .043$). Also, a significant difference was obtained between the slope for CAR less than 18 compared to when CAR is greater than or equal to 18 ($p = .027$).

Figure 7. The regression surface for predicting MAPA given CAR and CESD

7 Conclusions

Generally, robust methods are aimed at characterizing the typical response. When distributions are skewed, the 20% trimmed mean and M-estimators tend to have values close to the median. In contrast, the mean can be highly atypical. However, situations arise where comparing the tails of distributions can be important. For example, in the Well Elderly study, one goal was to compare the intervention group to a control in terms of depressive symptoms. But the bulk of the participants had relatively low CESD scores prior to intervention, suggesting that for the typical person there would be little improvement after intervention. Comparing medians found no significant difference, but a plot of the distributions indicated that differences in the right tails might exist. Indeed, comparing the upper quantiles yielded a significant difference. More precisely, among the more depressed individuals, intervention was found to be beneficial [45].

When considering the many new methods that have been derived, there is a seemingly natural question. Which method is best? But based on current technology, there is perhaps a better question: How many methods does it take to understand how groups compare or the association among the variables of interest? Different methods are sensitive to different features of the data and can provide different perspectives that have clinical importance. The method that maximizes power is unknown. But a strategy that is clearly unsatisfactory is to use classic methods for means and assume all is well. Standard rank-based methods reduce problems associated with outliers, but again more modern techniques have practical advantages.

Although the optimal regression method cannot be determined with certainty, there are some broad strategies that might be used. First, take advantage of smoothers. And when comparing groups, it can be helpful to plot estimates of the distributions. (The R functions akerd and g2plot are two good choices, which use kernel density estimators.) As was illustrated, checks for curvature can be crucial. Second, do not assume that testing assumptions justifies methods that assume normality or homoscedasticity. All indications are that generally, the safest way of knowing whether a more modern method makes a practical difference is to actually try it. There is, however, the issue of multiplicity: controlling the probability of one or more Type I errors when multiple tests are performed. One possibility is to choose some robust method for general use and if it fails to reject, consider an exploratory study where other techniques are used to analyze the data. One could then perform a confirmatory study. This, of course, would take some effort, but simultaneously it is important to not miss strong associations among the bulks of the data due to the statistical method that was used. Another strategy is to use a few well chosen techniques and control the probability of one or more Type I errors in some appropriate manner. For example, use some modern improvement on the Bonferroni method such as the method derived by Rom [34].

Finally, it is not being suggested that least squares methods be abandoned. It is being suggested, however, that heteroscedastic methods are generally superior to homoscedastic methods and that checks on the impact of removing leverage points can be highly valuable. Also, robust regression estimators and smoothers deserve serious consideration. We have the technology for getting a deeper and more accurate understanding of data. Software for applying these methods is readily available. All that remains is taking advantage of modern methods when analyzing data.

REFERENCES


