MODERN REGRESSION METHODS THAT CAN SUBSTANTIALLY INCREASE POWER AND PROVIDE A MORE ACCURATE UNDERSTANDING OF ASSOCIATIONS

Rand R. Wilcox
Dept of Psychology
University of Southern California
and
H. J. Keselman
Dept of Psychology
University of Manitoba

September 16, 2011

ABSTRACT

During the last half century hundreds of papers published in statistical journals have documented general conditions where reliance on least squares regression and Pearson's correlation can result missing even strong associations between variables. Moreover, highly misleading conclusions can be made, even when the sample size is large. There are, in fact, several fundamental concerns related to non-normality, outliers, heteroscedasticity, and curvature that can result in missing a strong association. Simultaneously, a vast array of new methods have been derived for effectively dealing with these concerns. The paper (1) reviews why least squares regression and classic inferential methods can fail, (2) provides an overview of the many modern strategies for dealing with known problems, including some recent advances, and (3) illustrates that modern robust methods can make a practical difference in our understanding of data. Included are some general recommendations regarding how modern methods might be used.

Keywords: Robust Methods; Non-normality; Heteroscedasticity; Outliers; Curvature; Well Elderly Study; Depression; Black Sheep Effect

1 Introduction

As is evident, least squares regression and Pearson's correlation play a central role in psychological research. From basic principles, classic inferential methods include three fundamental assumptions. The first is that the dependent variable has a normal distribution. The second is that there is homoscedasticity. That is the (conditional) variance of the dependent variable does not depend on the values of the independent variable(s). For example, if Y is some dependent variable of interest, such as the attitude of the participants, and X is some independent variable, say age, then the variance of Y does not depend on the age of the participants. So if the variance is 12 among children age 8, say, the variance is also 12 for children aged 7, 9, or any age of interest. If the variances change with age, there is heteroscedasticity. The third common assumption is that with one independent variable, the regression line is straight. More generally, with multiple independent variables, it is assumed the regression surface is a plane.

An issue of fundamental importance is whether there are general conditions, or features of data, that can result in poor power when testing the hypothesis of no association, or a misleading summary of the data when these assumptions are violated. Based on hundreds of papers published during the last half century, the answer is an unequivocal yes (e.g., Heritier et al., 2009; Huber & Ronchetti, 2009; Rousseeuw & Leroy, 1987; Staudte & Sheather, 1990; Wilcox, 2012, in press). From a purely descriptive point of view, least squares regression and Pearson's correlation can fail miserably, even when the sample size is large. A positive feature of classic inferential methods is that when there is independence, they control the probability of a Type I error reasonably well. If there is an association, again they might perform well in terms of power, but there are general conditions for which this is not the case. A practical implication is that when standard methods reject, it is reasonable to conclude that there is an association with the understanding that the usual linear model can poorly reflect the nature of the association. But when they fail to reject, concluding there is no association is not remotely justified. Indeed, even when least squares regression yields a highly non-significant result (a p-value substantially higher than .05), more modern methods can have a much higher probability of detecting a truly strong association. Also, as will be illustrated, there are important issues related to curvature that are not obvious based on standard training.

The paper is organized as follows. Section 2 reviews why violations of assumptions are a serious concern and it outlines the features of data that can result in poor power and misleading conclusions. Included is a summary of why the better-known methods for dealing with violations of assumptions perform poorly by modern standards. Section 3 briefly summarizes modern robust techniques for dealing with known problems. Virtually all of the concerns summarized in section 2 can be addressed in a very effective manner via modern techniques. Section 4 discusses curvature and section 5 describes software aimed at implementing robust regression methods. Section 6 provides more illustrations that these

newer methods can make a practical difference in our understanding of data. The paper concludes with some general suggestions about how to proceed in light of modern insights.

2 Practical Concerns with Least Squares Regression, Pearson's Correlation, and Standard Hypothesis Testing Techniques

This section discusses the practical consequences of violating each of the fundamental assumptions previously mentioned. Violating any of these assumptions is now known to be a serious concern. Not surprisingly, when two or more assumptions are violated simultaneously, as is often the case, problems are exacerbated. Features of data that can grossly distort the nature of the association among the bulk of the participants are discussed as well.

2.1 Outliers and Heavy-Tailed Distributions

One basic concern is the effect of outlying values. There is already some awareness of this issue among psychologists specializing in personality theory, where a trimmed mean has been used to address the negative effects of outlying values (e.g., Borkenau, 2010; Krause et al., 2011). The fact that outliers can seriously impact the least squares regression line, as well as Pearson's correlation, is now pointed out in some popular introductory texts (e.g., Moore & McCabe, 1999), but it seems apparent that the problem is under appreciated, possibly because most textbooks do not illustrate just how serious the problem can be. Moreover, the choice of method for detecting outliers can be crucial. Also, care must be taken regarding how to deal with outliers, when testing hypotheses, for reasons to be explained but which are not typically covered in most introductory books. There are general conditions where simply discarding outliers and testing hypotheses using the remaining data yields invalid results regardless of how large the sample size might be. There are theoretically sound methods for dealing with outliers (e.g., Heritier et al., 2009; Wilcox, 2012, in press), but they are not obvious based on standard training. In some situations, outliers can be beneficial, as will be explained. But they can lead to grossly inaccurate conclusions about the typical participants as well.

First focus on the situation where there are outliers among the dependent variable, Y. Figure 1 shows a scatterplot of data from an unpublished study dealing with predictors of reading ability among children. (The data were supplied by L. Doi.) The independent variable is taken to be a measure of phonological awareness and the dependent variable is the accuracy of identifying lower case letters. Pearson's correlation is -0.12 and the p-value based on the usual Student's t test is 0.29, suggesting there is little or no association.

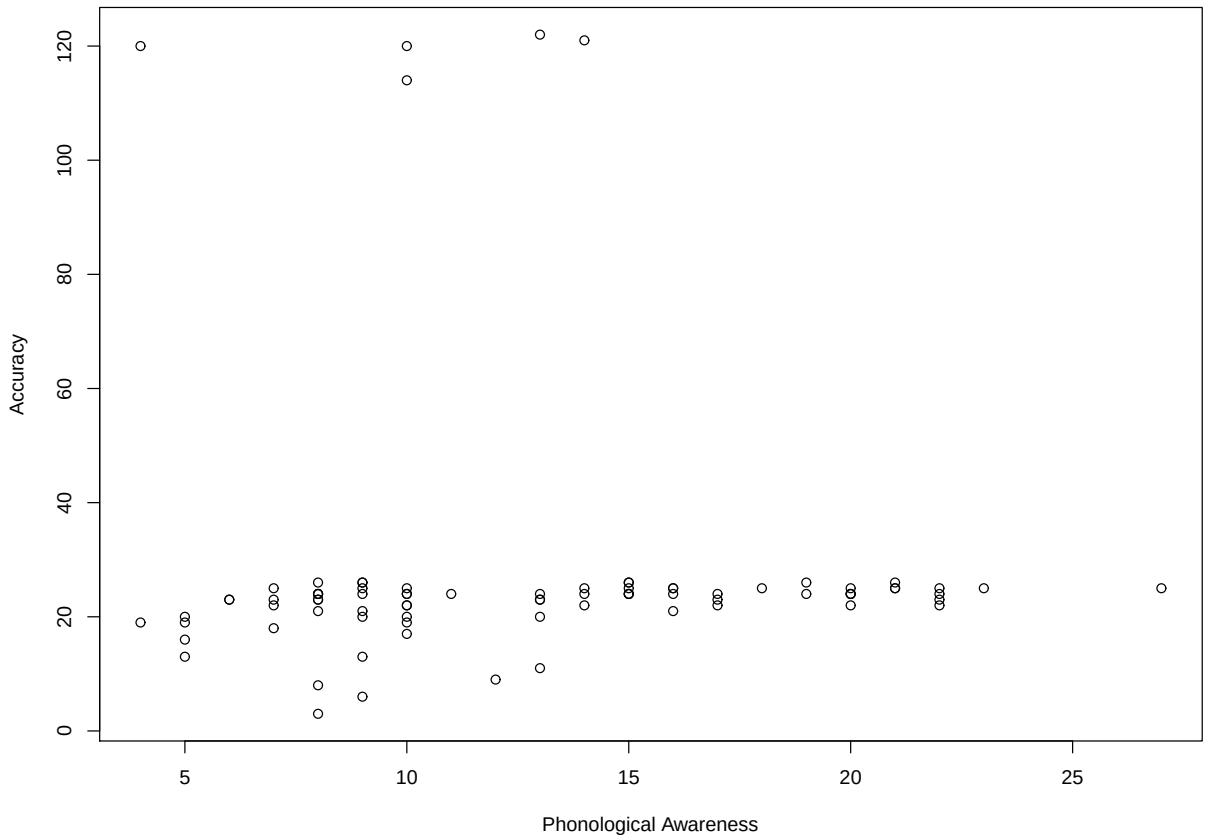

Figure 1: Outliers among the dependent variable, as shown here, can adversely affect power. Modern hypothesis testing techniques have been derived to handle this issue in a technically sound manner.

However, it is evident that the five largest accuracy scores are outliers and do not represent the typical accuracy scores. It might seem that one could simply remove these outliers and apply Student's t to the remaining data. But when using Student's t to test hypotheses based on the least squares regression estimate of slope, this results in a technical error: an incorrect estimate of the standard error is being used. This is because the standard error associated of the slope estimator is derived assuming that the observations are independent. The concern is that *if outliers are removed that are valid values, the remaining values are dependent.* In practical terms, *this can yield a highly inaccurate result regardless of how large the sample size might be.* This fact is well known in mathematical statistics, a simple explanation is given in Wilcox (2012, section 4.7.1), but it is evident that this fact is not well known otherwise. Moreover, and perhaps more importantly, *using technically sound methods for dealing with this problem can result in a substantially different conclusion.*

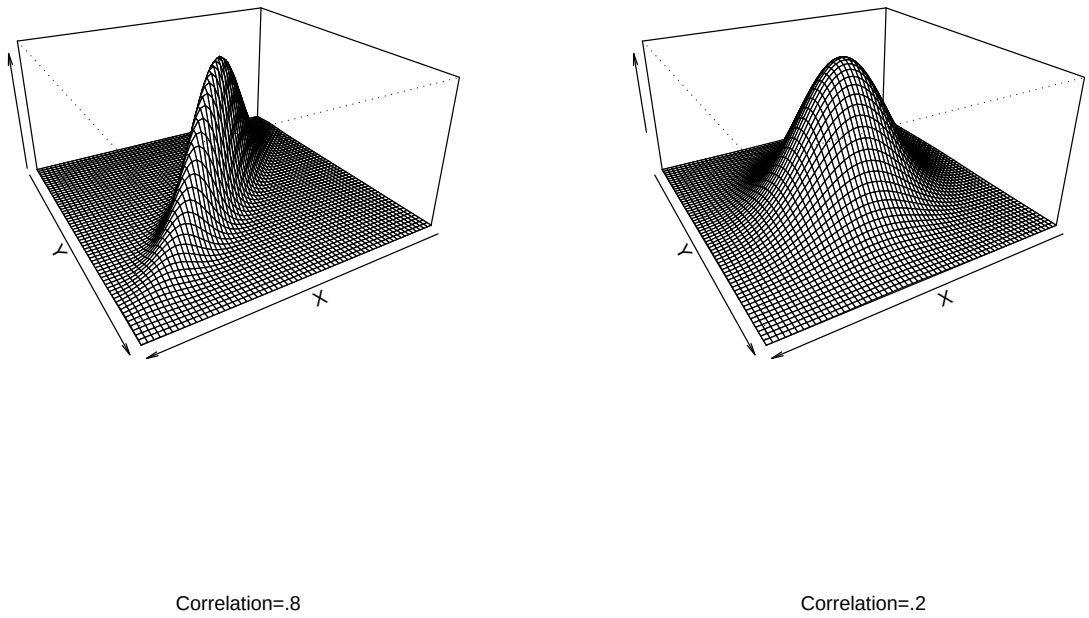

Figure 2: When both X and Y are normal, increasing ρ from .2 to .8 has a noticeable effect on the bivariate distribution of X and Y.

Pearson's correlation can miss a strong association even with a very large sample size. The left panel of Figure 2 shows a bivariate normal distribution having correlation .8 and the right panel shows a bivariate normal distribution having correlation .2. Now look at Figure 3. This bivariate distribution indicates a strong association from a graphical perspective, it has an obvious similarity with the left panel of Figure 2, yet Pearson's correlation is only .2. The reason is that one of the marginal distributions has what is called a mixed normal distribution, which is an example of a heavy-tailed distribution, roughly meaning that outliers are more common compared to when data are normally distributed. This illustrates the fact that *a small change in the tails of one of the distributions can have a large impact on Pearson's correlation.* Robust measures of association are designed to guard against this event.

Numerous methods are now available for dealing with outliers, among the dependent variable, in a technically sound manner (Wilcox, 2012, in press). A simple class of methods

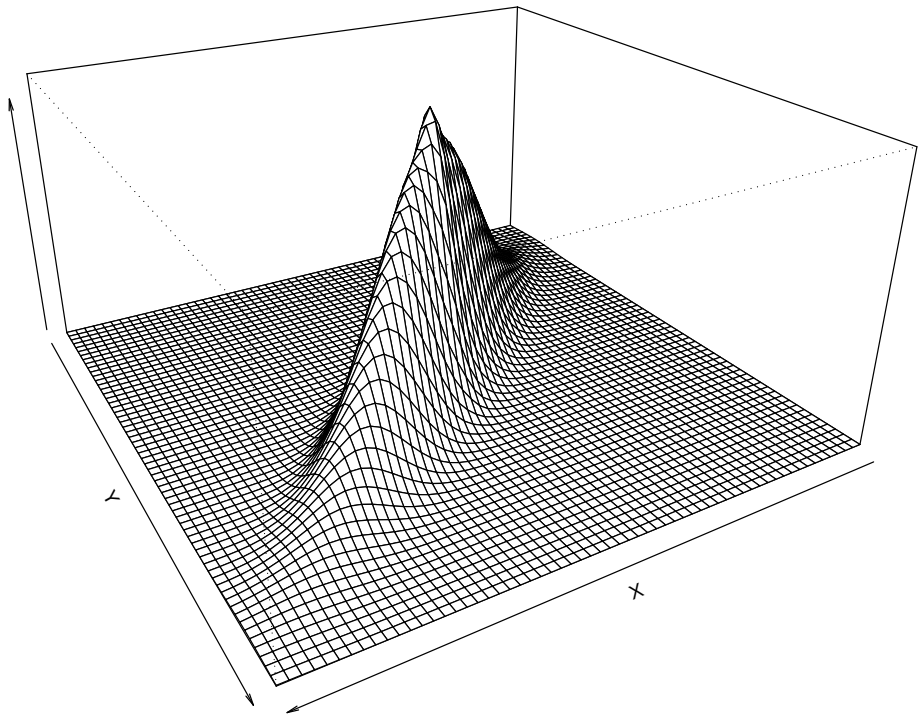

Figure 3: Two bivariate distributions can appear to be very similar yet have substantially different correlations. Shown is a bivariate distribution with $\rho = .2$, but the graph is very similar to the left panel of Figure 2 where $\rho = .8$.

is based on the outlier detection methods outlined in section 2.2. Typically, robust measures of association and regression methods are designed to give similar results to least squares regression and Pearson's correlation when standard assumptions are true. But an additional goal is to avoid low power and misleading conclusions when standard assumptions are false. Here it is merely remarked that for the data in Figure 1, a 20% Winsorized correlation is equal to .34, and the resulting p-value is .002. The so-called OP correlation, which detects outliers using a projection method, is .42 with a p-value of .003. *So robust measures of association can differ substantially from Pearson's correlation.* (The distinction between these two robust correlations is discussed in section 3.2.)

Outliers among the independent variable are called leverage points. There are two kinds: good and bad, which are illustrated in Figure 4. Roughly, a good leverage point is one that is consistent with the regression line associated with the bulk of the data. One positive appeal of a good leverage point is that it results in a smaller standard error compared to situations where there are no leverage points. A bad leverage point has the potential to substantially affect the least squares regression line in a manner that completely distorts the association among the majority of the participants. This is illustrated in Figure 5. The left panel deals with data on the average influent nitrogen concentration (NIN) in 29 lakes and the mean annual total nitrogen (TN) concentration. Note the obvious leverage points in the lower right portion of the plot. The nearly horizontal (solid) line is the least squares regression line using all of the data. Based on the usual Student's t test, the p-value associated with the slope is .72. The other regression (dotted) line is the least squares regression line when the leverage points are removed. The right panel is from the reading study, only now a measure of digit naming speed (RAN1T) is used to predict the ability to identify words (WWISST2). The nearly horizontal (solid) line is the least squares regression line using all of the data and the other (dotted) line is the regression line when the obvious leverage points are removed.

From an inferential point of view, dealing with leverage points is straightforward. Removing outliers among the independent variable does not invalidate the derivation of the standard error. For the situations in the left panel of Figure 5, eliminating the obvious leverage points, the p-value is .0003. For the reading study, the p-value associated with the slope, using all of the data, is .76. Eliminating the six obvious leverage points, the p-value drops to .002.

2.2 Detecting Outliers

A point worth stressing is that the choice of method for detecting outliers can be crucial. One can quibble about the best method, but there is no controversy over the fact that a highly unsatisfactory approach (due to masking) is any method based on the usual mean and variance. For example, declaring a point an outlier if it is more than two standard deviations from the mean is highly unsatisfactory due to masking. That is, *the very presence*

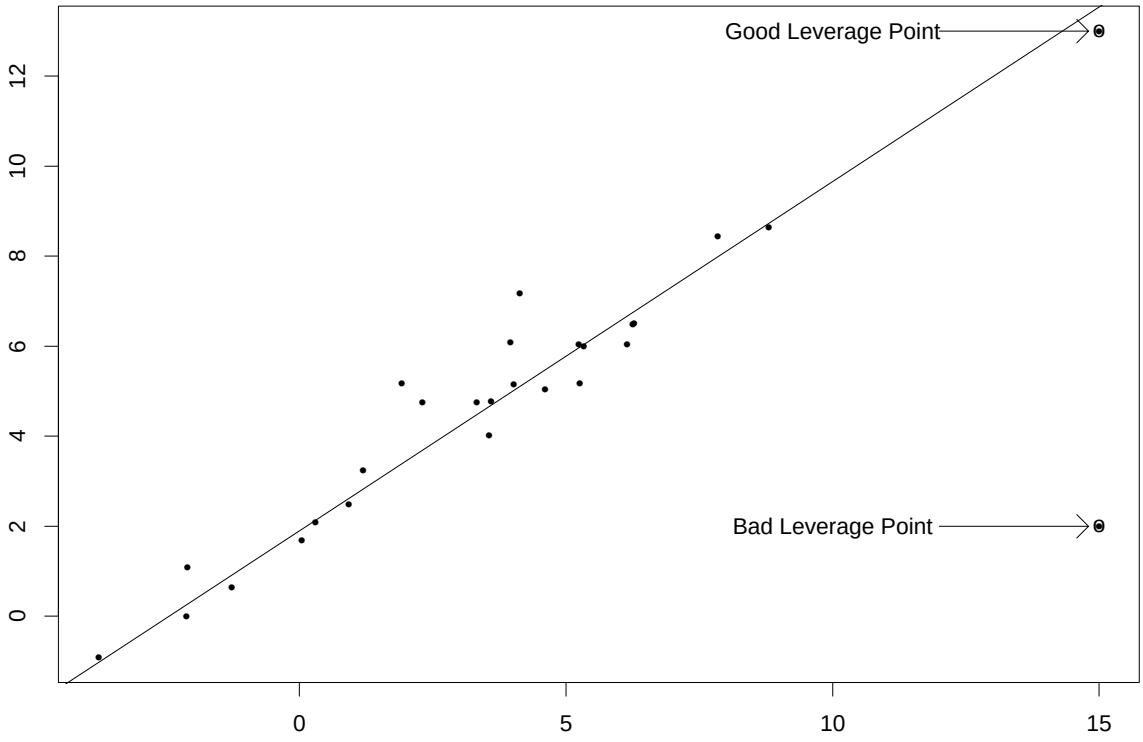

Figure 4: An illustration of good and bad leverage points.

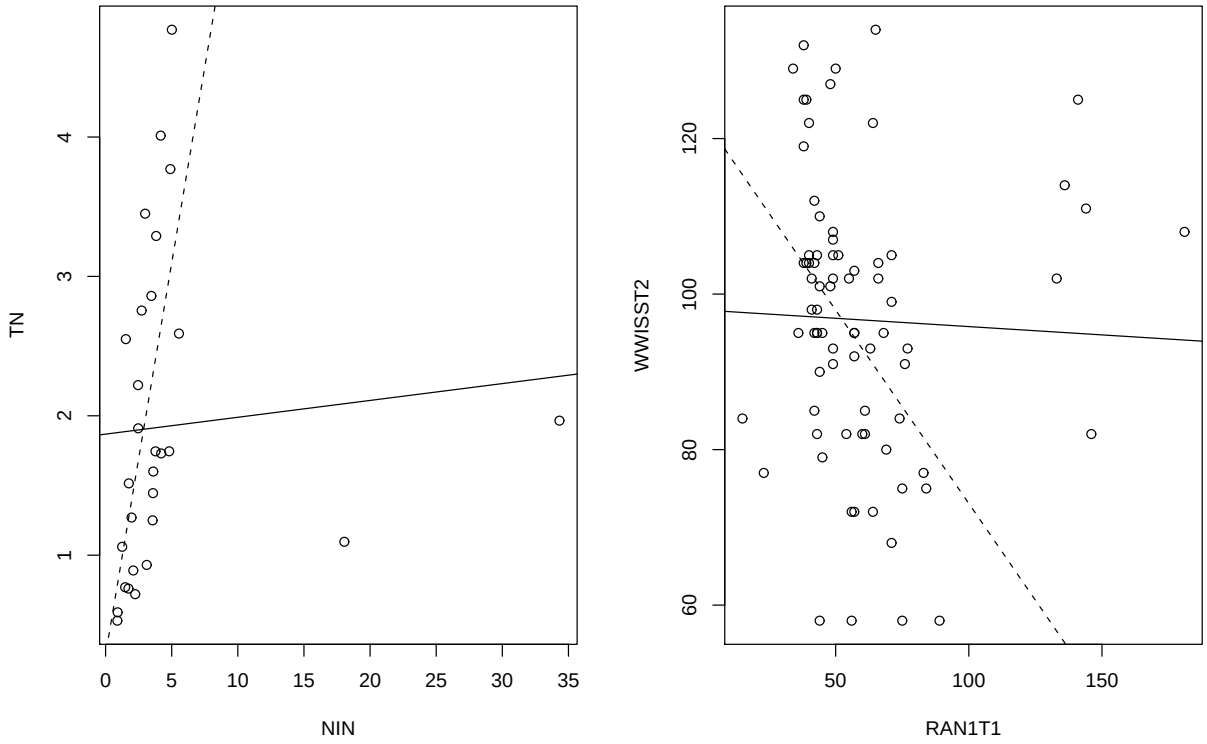

Figure 5: Removing leverage points, using a good outlier detection method, can have a large impact on the estimated regression line and the p-value.

of outliers can cause all outliers to be missed, regardless of how extreme they might be. The reason is that although the mean is sensitive to outliers, the standard deviation is even more sensitive to outliers in the sense that it increases faster, as an outlier becomes more extreme, than the difference between the outlier and the sample mean. More satisfactory methods are the boxplot rule and the so-called MAD-median rule. This latter method is applied as follows. Let M be the usual median based on the observations $X_1, \ldots, X_n$. The median absolute deviation (MAD) is the median of the values $|X_1 - M|, \ldots, |X_n - M|$. The value X_i is declared an outlier if

$$\frac{|X_i - M|}{MAD/.6745} > 2.24.$$

Multivariate data requires more involved techniques. Simply checking for outliers among the marginal distributions can be unsatisfactory. What is needed is a method that takes into account the overall structure of the data. For example, it is not unusual to be young, and it is not unusual to have hardening of the arteries, but it is unusual to be both young and have hardening of the arteries. A method based on the usual covariance matrix (Mahalanobis distance) is highly unsatisfactory due to masking. Several methods that deal effectively with multivariate data are available and described in Wilcox (in press, section 6.4). One general approach replaces the usual mean and covariance matrix with a robust analog (many of which have been proposed). A fast and popular choice is the minimum volume ellipsoid estimator. A criticism, however, is that this method might declare too many points outliers when sampling from a multivariate normal distribution. A better choice is a so-called projection method. Both of these methods, plus several others, can be applied with the software described in section 5.

2.3 Heteroscedasticity

There are at least two serious concerns associated with heteroscedasticity. Let b_1 be the estimate of the slope. Recall from basic principles that when testing the hypothesis that a slope parameter is equal to zero, the test statistic is based in part on an estimate of the standard error of b_1. But the expression for the standard error is based on the assumption that there is homoscedasticity. *Violating this assumption can result in an actual Type I error probability exceeding .5 when testing at the .05 level.* A positive feature of this approach is that independence implies homoscedasticity. That is, if there is no association, a reasonable expression for the standard error is being used. But when there is an association, there is no reason to assume homoscedasticity. *If there is heteroscedasticity, an incorrect estimate of the standard error is being used that can affect power and the accuracy of any confidence interval. In fact, the classic estimate of the standard error can differ from a theoretically sound estimate (which allows heteroscedasticity) by a factor of two. Also, the usual Student's t test that Pearson's correlation is equal to zero, assumes homoscedasticity, which again can result in poor power and misleading conclusions.*

A seemingly natural strategy is to test the assumption that there is homoscedasticity. Numerous methods for testing this assumption have been proposed, the majority of which have been found to be unsatisfactory in terms of controlling the probability of a Type I error. Two that control the probability of a Type I error reasonably well are available (e.g., Wilcox, in press, section 11.3). But a concern about these tests is that they might not have enough power to detect situations where the violation of the homoscedasticity assumption results in inaccurate results (e.g., Ng & Wilcox, 2011). *A better approach is to use a method that allows heteroscedasticity.* Many such methods are now available that perform about as well as homoscedastic methods when the homoscedasticity assumption is true. (A possible argument for testing the hypothesis of homoscedasticity is that it provides a way of detecting a particular type of dependence that is not detected by other techniques.)

When using least squares regression, methods that estimate a correct expression for the standard error are available (e.g., Long & Ervin, 2000; Godfrey, 2006; Cribari-Neto, 2004). Two approaches are based on what is called the HC4 estimate of the standard error, one of which is based in part on a particular bootstrap technique. Extensive simulations indicate that these methods substantially improve upon homoscedastic methods, but there are situations where control over the probability of a Type I error remains unsatisfactory (Ng & Wilcox, 2006; Wilcox, 2012, section 6.4.6), even with a sample size of $n = 100$. Yet another concern is that *compared to several modern robust regression estimators, the least squares estimator can have a relatively high standard error, which can result in relatively low power. Robust regression estimators, coupled with improved hypothesis testing techniques, deal effectively with this issue. That is, there can be practical advantages to replacing least squares regression with one of the robust estimators in section 3.* Moreover, hypotheses can be tested using a percentile bootstrap method (outlined in section 3.1), which deals effectively with heteroscedasticity.

3 Robust Regression Methods

There are numerous robust regression estimators designed to deal with outliers. Currently, the most comprehensive summary, including a discussion of their relative merits, is given in Wilcox (in press, chapter 9). Although space limitations prohibit a thorough summary of all the estimators available, an outline of the strategy behind a few of the estimators should reveal their operating characteristics.

Momentarily consider the situation where there is one independent variable only. The Theil (1950) and Sen (1968) estimator is applied as follows. Compute the slope of the line between each pair of distinct points. The median of all such slopes is the Theil–Sen estimate of the slope, say b_{ts}. The intercept is taken to be $M_y - b_{\text{ts}}M_x$, where M_y and M_x are the median of the dependent and independent variables, respectively. There are several ways of extending this estimator to situations where there is more than one independent variable

(Wilcox, in press, section 10.2), but the details are not important here.

Another general approach is to determine the slopes and intercept that minimizes some robust measure of variation applied to the residuals. There are many robust scale estimators (measures of variation), comparisons of which have been made by Lax (1985) and Randal (2008). The so-called percentage bend midvariance performs relatively well. A somewhat related approach is least trimmed squares where the goal is to determine the slopes and intercepts so as to minimize the sum of the squared residuals, ignoring some specified proportion of the largest residuals. (This strategy has some similarities to a trimmed mean.) Yet another general approach is based on what are called M-estimators. *A crude description is that M-estimators empirically downweight or eliminate unusually large or small residuals.* (The formal definition of an M-estimator is more involved.) Two that perform relatively well are the Coakley and Hettmansperger (1993) estimator and the MM-estimator derived by Yohai (1987). *There are conditions where all of these estimators offer a substantial advantage over least squares regression in terms of power, control over the probability of a Type I error probability, and the ability to avoid misleading summaries of data due to outliers.* When standard assumptions are true, least squares does not offer much of an advantage. However, no single regression estimator is always best. The Theil–Sen estimator and the MM-estimator seem to perform relatively well but there are practical reasons for considering other estimators that are difficult to explain quickly. (See Wilcox, in press, chapter 10.)

3.1 Hypothesis Testing

For some robust regression estimators, expressions for the standard errors are available. But hypothesis testing methods based on estimates of the standard error are not recommended for two reasons. First, homoscedasticity is assumed. Second, test statistics that make use of these standard errors can perform poorly when dealing with skewed distributions. (This is generally true for M-estimators.) A better approach is to use a percentile bootstrap method, which performs well, in terms of Type I errors, even when there is heteroscedasticity and when distributions are skewed. The hypothesis that all slope parameters are equal to zero can be tested using the software described in section 5, but the involved computational details are not given here. However, to at least impart the flavor of the method, it is described when the goal is to test the hypothesis that a slope parameter is zero.

Suppose the goal is to test the hypothesis that the slope parameter is equal to zero. A percentile bootstrap begins by resampling with replacement, n vectors of observations from the observed data having sample size n. Compute the estimate of the slope and label it b^*. Repeat this process B times, determine the proportion of times b^* is greater than the hypothesized value, and label the result P. (In terms of controlling the Type I error probability, $B = 600$ appears to suffice.) A p-value is given by

$$2\min(P, 1 - P)$$

(Liu & Singh, 1997). A .95 confidence interval corresponds to the middle 95% of the b^* values after the values are put in ascending order. (A global test that all slope parameters are equal to zero is available as well.)

The method just described appears to perform very well when using a robust regression estimator designed to deal with outliers. But when using least squares regression, alternative methods should be used that seem a bit more satisfactory (Wilcox, 2012, sections 6.4.6 and 7.7).

3.2 Measures of Association

A simple robust analog of Pearson's correlation is the Winsorized correlation. Recall that trimming 20% means that the 20% smallest and largest values are removed. Winsorizing data means that rather than remove the smallest observations, they are set equal to the smallest value not trimmed. Similarly, the largest 20% of the observations are set equal to the largest value not trimmed. Like Spearman's rho and Kendall's tau, *the Winsorized correlation guards against outliers among the marginal distributions*, but it does not take into account the overall structure of the data. A method that does is the OP measure of correlation, which is based in part on a projection method for detecting outliers. Simple methods for testing the hypothesis of no association are available, assuming homoscedasticity. To avoid sensitivity to heteroscedasticity, a percentile bootstrap method can be used.

Another general approach to measuring association is explanatory power, which has the advantage of being readily applied when dealing with curvature. For a set of predictors, $X_1, \ldots X_n$, let $\hat{Y}_i$ be the predicted value of the dependent variable based on X_i, the ith vector of observations, and some regression model. Explanatory power is the variance of the predicted Y values, the $\hat{Y}_i$ values, divided by the variance of the observed Y values. A robust version is obtained simply by replacing the variance with some robust measure of scatter. One of the better choices is the percentage bend midvariance, but arguments for considering other measures of variation can be made. If $\hat{Y}_i$ values are based on the least squares regression estimator, and the usual variance is used, explanatory power reduces to r^2, the squared Pearson correlation. *The square root of explanatory power is called the explanatory measure of association.*

4 Curvature

Modern technology offers an array of tools for dealing with curvature that can make a substantial difference in our understanding of data, including the goal of determining whether an association exists (e.g. Efromovich, 1999; Eubank, 1999; Fan and Gijbels, 1996; Fox, 2001;

Green and Silverman, 1993; Gyofri et al., 2002; Härdle, 1990; Hastie and Tibshirani, 1990; Wilcox, 2012, in press). Of particular importance are *smoothers*, which provide *a flexible approach to approximating a regression surface*. Experience with smoothers makes it clear that the more obvious classical parametric models for dealing with curvature can be highly unsatisfactory, particularly when dealing with more than one predictor. Some illustrations supporting this statement are given in section 6. *Modern methods for dealing with curvature can reveal strong associations that otherwise would be missed using the more traditional techniques*, as will be seen.

Imagine that the typical outcome of some dependent variable is given by some unknown function $m(X_1, \ldots, X_p)$ of the p predictors $(X_1, \ldots, X_p)$. A crude description of smoothing techniques is that they identify which points, among n vectors of observations, are close to $(X_1, \ldots, X_p)$, and then some measure of location is computed based on the corresponding Y values. The result is $\hat{m}(X_1, \ldots, X_p)$, an estimate of the measure of location associated with Y at the point $(x_1, \ldots, x_p)$. The estimator $\hat{m}$ is called a *smoother*, and the outcome of a smoothing procedure is called a *smooth* (Tukey, 1977). A slightly more precise description of *smoothers* is that they *are weighted averages of the Y values with the weights a function of how close the vector of predictor values is to the point of interest*. In more formal terms, one estimates $m(x_1, \ldots, x_p)$ with

$$\hat{m}(x_1, \ldots, x_p) = \sum w_i y_i, \tag{1}$$

and the goal is to choose the w_i values in some reasonable manner. Many suggestions regarding the choice of weights have been made. The illustrations used here are based on methods derived by Cleveland (1979) as well as Cleveland and Devlin (1988).

Note that for the case of a single independent variable, an estimate of Y can be computed for each observed X_i value $(i = 1, \ldots n)$. Plotting the estimates versus the X_i values provides a graphical indication of the regression line. (Various refinements of this strategy have been derived.) In a similar manner, when there are two independent variables, again a graphical representation of the regression surface can be created. Such graphs can be invaluable when trying to detect and describe an association. Moreover, there are inferential methods based on these smoothers. For example, one can test the hypothesis that a regression line is straight or that it has a particular parametric form. *Smoothers provide a much more flexible approach to comparing groups when there is a covariate*. These modern ANCOVA methods effectively deal with non-normality and heteroscedasticity as well. Moreover, moderator analyses can be performed that can greatly enhance the probability a detecting a true interaction when there is curvature.

Generally, smoothers can be used to compute predicted Y values for each observed X value. These predicted Y values can be used in turn to compute a robust measure of explanatory power.

5 Software

Of course, modern robust methods are useless without access to appropriate software. The free software R, which can be downloaded from www.R-project.org, is easily the best software for dealing with this issue. The power and flexibility of R help explain why it dominates among statistics books written by statisticians, as well as why it is beginning to be replace SPSS among some psychologists. (SPSS is hopelessly antiquated when it comes to taking advantage of modern technology.) All of the illustrations in this paper are based on R in conjunction with the R functions described and illustrated in Wilcox (2012, in press). Wilcox's R package contains over 950 functions for applying modern methods.

Using R to apply robust regression methods is straightforward. For example, if the data for one or more predictors are stored in the R variable x, and the dependent variable is stored in y, the R command tsreg(x,y) computes the Theil–Sen regression estimator. The command tsreg(x,y,xout=T,outfun=outpro) computes the Theil–Sen estimator after removing leverage points via a projection method for multivariate data. The functions regci and regtest test hypotheses. By default, the Theil–Sen estimator is used, but any robust regression estimator can be used via the argument regfun. The functions lplot, lintest, and adtest plot an estimate of the regression surface, test the hypothesis that a linear model fits the data, and tests the hypothesis that there is no interaction, respectively. The function scor computes the OP measure of association.

6 More Illustrations

The first two illustrations stem from an unpublished study by G. Stratton, N. Miller and B. Lickel dealing with the black sheep effect, which refers to an exaggerated harsh response to a negatively deviant ingroup member by comparison with a similarly deviant outgroup member. A portion of the study dealt with the association between a measure of effect size (Cohen's d) from 99 studies and various independent variables. One of these variables, labeled expectancies of the outgroup, was found to have a Pearson correlation of $-.15$ with a p-value of .15 based on the usual Student's t test. However, look at the regression line shown in Figure 6 based on a smoother. (The R function lplot was used.) It suggests that initially there is a negative association, but at some point the association virtually disappears. The explanatory strength of association is .21. The hypothesis that the regression line is straight is rejected at the .028 level. That is, the data indicate that there is an association, but it is not well represented by a straight line. *Experience with smoothers suggests that it is common to encounter situations where there is an association over some range of an independent variable, but that otherwise there is little or no association.*

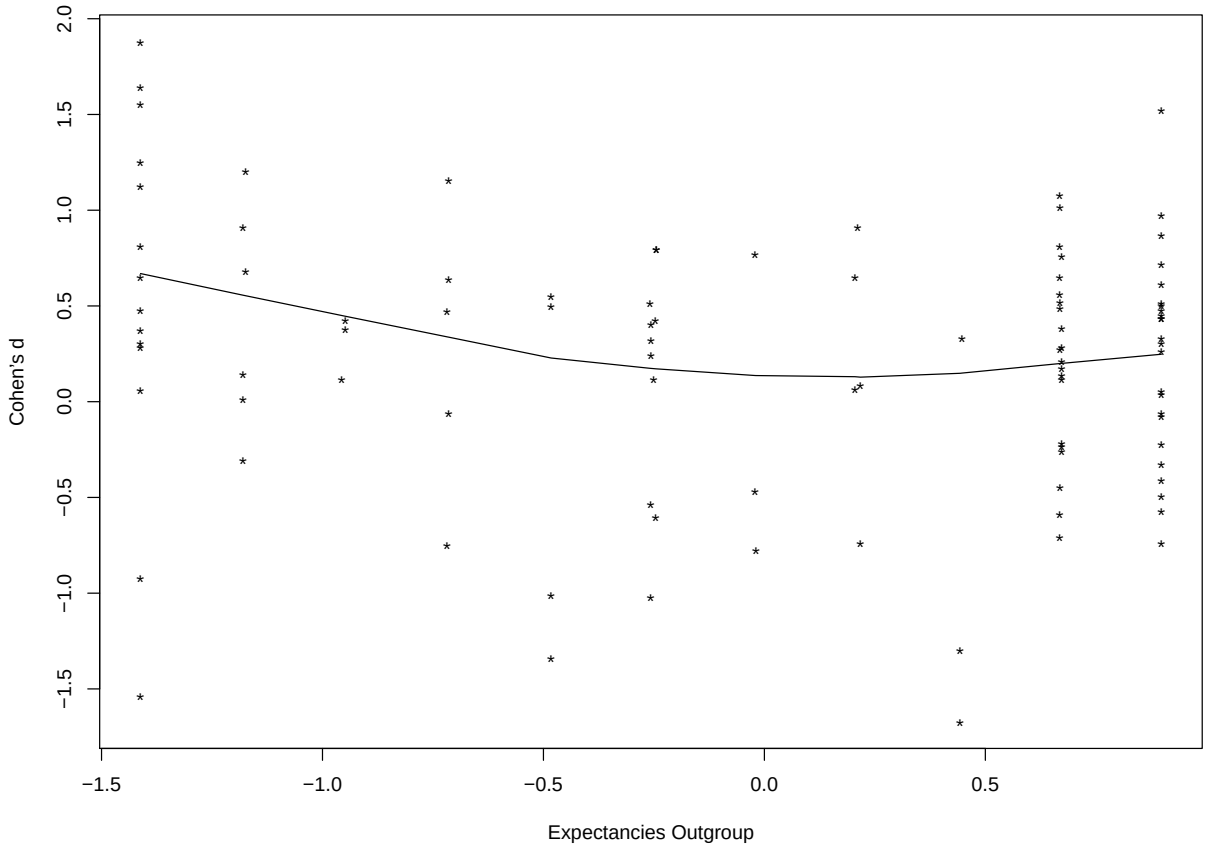

Figure 6: It is common to encounter situations, such as shown here, where there is an association over some range of the independent variable but otherwise no association is found.

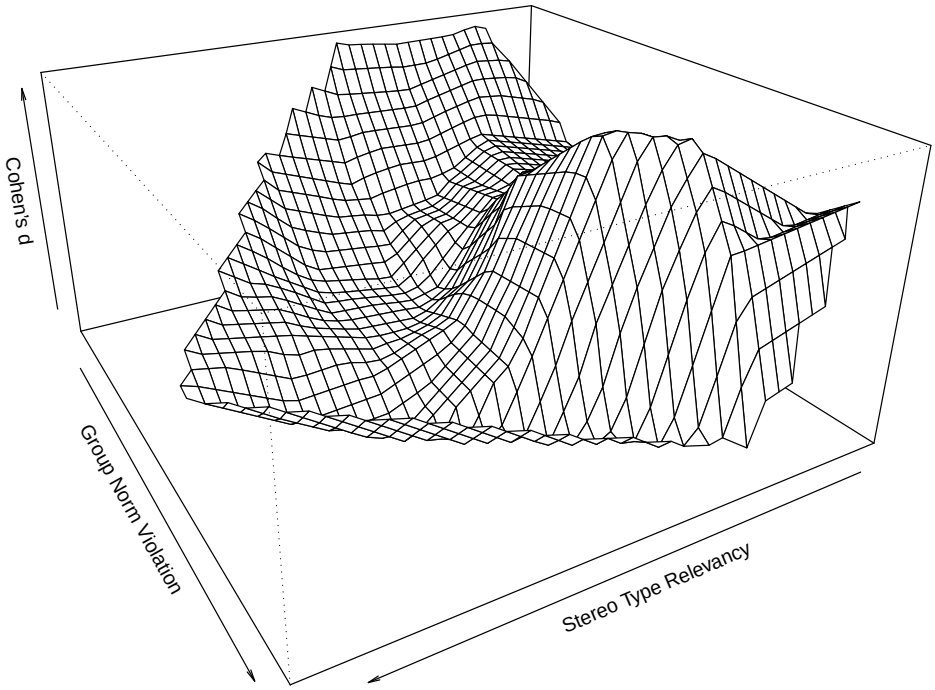

Figure 7: Curvature can be an issue with two predictors, even when it is not an issue for the individual predictors

When dealing with more than one independent variable, taking into account curvature can make a more striking difference. Again consider the Stratton et al. study, only now focus on the variables group norm violation and stereo type relevancy. The Pearson correlations with Cohen's d are .03 and .05, respectively. The corresponding Student's t test p-values are .78 and .63. A smooth for each of the regression lines appears to be reasonably straight and horizontal. But this is not compelling evidence that, when taking together, there is no curvature (e.g., Berk & Booth, 1995). Figure 7 shows a smooth of the regression surface when both predictors are used simultaneously. (Leverage points were removed using the MVE method.) The robust explanatory measure of association is .46 and the hypothesis that the regression surface is a plane is rejected at the .04 level.

In another unpublished study by S. Tom and D. Schwartz, one goal was to understand the association between a Totagg score and two predictors: grade point average (GPA) and a measure of academic engagement. The Totagg score was a sum of peer nomination items

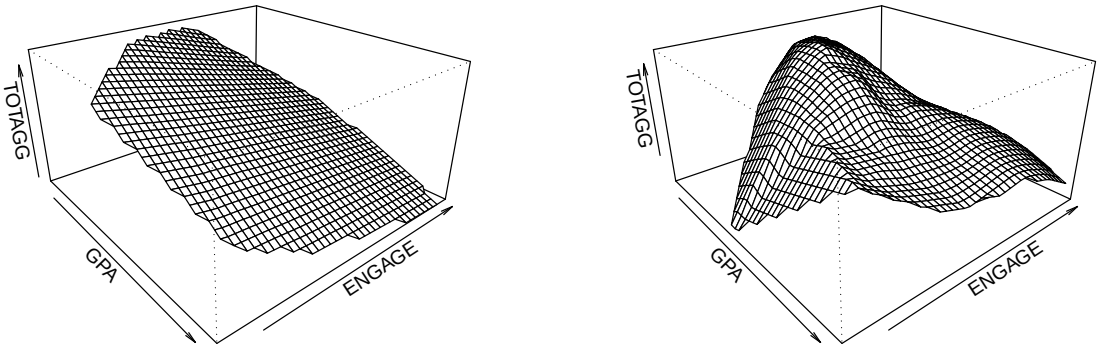

Figure 8: The left panel shows the regression surface based on the usual interaction model, which uses the product of the two predictors. The right panel shows an approximation of the regression surface using a smoother, which provides a more flexible approach to dealing with curvature.

that were based on an inventory that included descriptors focusing on adolescents' behaviors and social standing. (The peer nomination items were obtained by giving children a roster sheet and asking them to nominate a certain amount of peers who fit particular behavioral descriptors.) The sample size was $n = 336$. One issue is whether there is an interaction between these two predictors. If we use least squares regression and the usual product term, the p-value is .64 using the HC4 method for dealing with heteroscedasticity. The left panel of Figure 8 shows a plot of the regression surface based on this standard approach. (The plot was created with the R function ols.plot.inter.) The right panel shows an approximation of the regression surface using a smoother. (The R function lplot was used.) This more flexible approach to approximating the regression surface differs in obvious ways from the left panel. Testing the hypothesis of no interaction (using the R function adtest), the p-value is less than .01.

The next example stems from the Well Elderly study (Clark et al., 1997). One of the many goals was to understand the association between depression and two biomarkers: dehydroepiandrosterone (DHEA) and cortisol, both of which are produced by the adrenal glands and known to be associated with stress and depression. Figure 9 shows an estimate of the regression surface when predicting depression as measured by the Center for Epidemiologic Studies Depression Scale (CESD), and when DHEA and cortisol are measured upon awakening. (The CESD was developed by Radloff, 1977.) The estimated strength of the association is .25. (Leverage points were identified using a projection method and then removed.) The plot indicates that for low cortisol, as DHEA increases, depression decreases up to a point and then levels off. However, when cortisol is relatively high, depression is relatively high even when DHEA is high. The hypothesis that the regression surface is a plane is rejected, $p < .002$. The hypothesis of no interaction is rejected as well, $p < .002$. In contrast, no association is found using the usual F test associated with the least squares estimator, and no association is found using the Theil–Sen estimator.

7 Some Suggestions Regarding How To Proceed

Different methods are sensitive to different features of the data. *No single method is always optimal.* But some general guidelines regarding how to proceed might help:

- If the goal is to make inferences about the slope parameters using a linear model, use a method that allows heteroscedasticity.
- If a homoscedastic method is used, interpret it properly: if it rejects, there is dependence. But the main reason for rejecting might be due to heteroscedasticity. Never conclude there is no association based on a homoscedastic method.
- Consider what happens when leverage points are removed. (This is easily done with the R functions used here by setting the argument xout=T.)
- Consider a robust regression estimator that deals effectively with outliers among the dependent variable.
- Consider methods that provide a flexible approach to curvature. In particular, use regression smoothers as a partial check on the nature of the association. Never assume that curvature can be ignored. With one or two independent variables, plot a smooth of the regression surface. (The R function lintest can be used to test the hypothesis that a particular parametric model fits the data.)

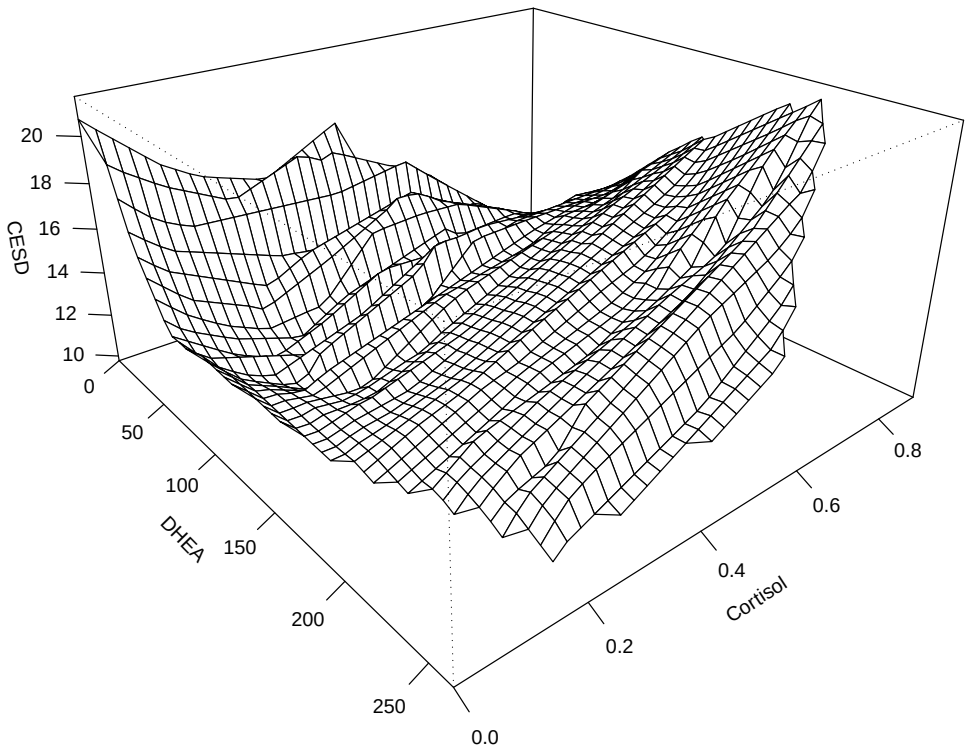

Figure 9: Regression surface when predicting CESD with DHEA and Cortisol upon awakening.

8 Concluding Remarks

Vast sums of money and time are spent on collecting data relevant to a wide range of important and interesting issues. *The weak link in this process is clear: data analyses. Modern technology offers a variety of improved tools for discovering associations and describing them in a more accurate and informative manner.*

Modern methods offer advantages that go beyond the scope of this paper. For example, many introductory statistics books still claim that with a sample size of 30 or more, normality can be assumed. There are reasons for this claim, but two issues were missed when studying the properties of the central limit theorem. If attention is restricted to standard inferential methods, there are general conditions where even a sample size of 300 does not suffice. There are techniques for dealing with this issue, they can make a practical difference in our understanding of data, so it is suggested that they be considered when analyzing data.

To add perspective, consider all of the methods that are routinely taught and used, including all ANOVA designs and classic nonparametric methods. Generally, these methods perform well, in terms of Type I error probabilities when comparing groups that have identical distributions or when testing hypotheses about associations when in fact there is no association. When groups differ, or when there is an association, of course they might continue to perform well. But for a broad range of situations, this is not the case. Techniques for dealing with known problems are now available. Although some researchers are aware of modern methods and take advantage of them, it is evident that generally this is not the case. This raises an important issue: *How many discoveries are lost by ignoring modern methods*? All indications are that the answer is more than what would be expected based on standard training.

References

Berk,K. N. & Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics, 37*, 385–398.

Borkenau, P., Paelecke, M. & Yu, R. (2010). Personality and lexical decision times for evaluative words. *European Journal of Personality, 24*, 123-136.

Clark, F., Azen, S. P., Zemke, R., et al., (1997). Occupational therapy for independent-living older adults. A randomized controlled trial. *Journal of the American Medical Association, 278*, 1321–1326.

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

Cleveland, W.S., and Devlin, S.J., (1988). Locally-weighted Regression:

An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association, 83*, 596–610.

Coakley, C. W. & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association, 88*, 872–880.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications.* New York: Springer-Verlag.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* New York: Marcel Dekker.

Fan, J. & Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Boca Raton, FL: CRC Press.

Fox, J. (2001). *Multiple and Generalized Nonparametric Regression.* Thousands Oaks, CA: Sage.

Godfrey, L. G. (2006). Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics & Data Analysis, 50*, 2715–2733.

Green, P. J. & Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* Boca Raton, FL: CRC Press.

Györfi, L., Kohler, M., Krzyzk, A. & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression.* New York: Springer Verlag.

Härdle, W. (1990). *Applied Nonparametric Regression.* Econometric Society Monographs No. 19, Cambridge, UK: Cambridge University Press.

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models.* New York: Chapman and Hall.

Heritier, S., Cantoni, E, Copt, S. Victoria-Feser, M.-P., 2009. *Robust Methods in Biostatistics.* New York: Wiley.

Huber, P. J. & Ronchetti, E. (2009). *Robust Statistics*, 2nd Ed. New York: Wiley.

Koenker, R., 1981. A note on studentizing a test for heteroscedasticity *Journal of Econometrics, 17*, 107–112.

Krause, S., Back, M. D., Egloff, B. & Schmukle, S. C. (2011). Reliability of implicit self-esteem measures revisited. *European Journal of Personality, 25*, 239–251.

Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association, 80*, 736–741.

Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association, 92*, 266–277.

Long, J. S. & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician, 54*, 217–224.

Ng, M. & Wilcox, R. R. (2009). Level robust methods based on the least squares regression line. *Journal of Modern and Applied Statistical Methods, 8*, 384–395.

Ng, M., Wilcox, R. R. (2011). A comparison of two-stage procedures for testing least-squares coefficients under heteroscedasticity. *British Journal of Mathematical & Statistical Psychology, 64*, 244–258.

R Development Core Team, 2010. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Radloff L. (1977). The CES-D scale: a self report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385–401.

Randal, J. A. (2008). A reinvestigation of robust scale estimation in finite samples. *Computational Statistics & Data Analysis, 52*, 5014–5021.

Rasmussen, J. L. (1989). Data transformation, Type I error rate and power. *British Journal of Mathematical & Statistical Psychology, 42*, 203–211.

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77*, 663–666.

Rousseeuw, P. J., Leroy, A. M. (1987). *Robust Regression & Outlier Detection*. New York: Wiley.

Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63*, 1379–1389.

Staudte, R. G., Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae 12*, 85–91.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Wilcox, R. R. (2012). *Modern Statistics for the Social and Behavioral Sciences: A Practical Introduction*. New York: Chapman & Hall/CRC press

Wilcox, R. R. (in press). *Introduction to Robust Estimation and Hypothesis Testing*, 3rd Ed. San Diego, CA: Academic Press.

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics, 15*, 642–656.