

# Modern Insights About Pearson's Correlation and Least Squares Regression

Rand R. Wilcox\*

As is well known, Pearson's correlation,  $\rho$ , can be used to characterize how well a least squares regression line fits data, and it provides a test of the hypothesis that two measures are independent. However, many articles in statistical journals indicate that the usual estimate of  $\rho$ ,  $r$ , is sensitive to at least six features of data, and that least squares regression and  $\rho$  are not robust in the sense reviewed in this article. In practical terms,  $r$  can be a highly unsatisfactory measure of the strength of an association, no matter how large the sample size might be. One specific problem is that it can miss strong associations that are detected by more modern techniques. The practical problems with  $r$  reflect fundamental concerns about a strict reliance on least squares regression. A few of the many modern methods for dealing with these concerns are briefly indicated.

## Introduction

As is evident, Pearson's correlation,  $\rho$ , plays an integral role in personnel selection, and of course it is fundamental when assessing reliability and predictive validity. Often Pearson's correlation is the only tool used by psychologists to assess the association between two variables. There are some well-known interpretations about the magnitude of  $\rho$  which certainly have practical value. However, in recent years, a collection of insights and advances make it evident that Pearson's correlation might miss or underestimate strong associations. Indeed, arbitrarily small departures from normality can wreak havoc. The magnitude of Pearson's correlation tells us something about how well a least squares line fits data, regardless of whether normality is true, but hundreds of articles in statistical journals have pointed out that when using least squares regression, a single unusual point can give a highly distorted view of how the bulk of the observations are related. In fact, there are at least six features of data that influence the magnitude of  $r$ , the usual estimate of  $\rho$ , which are summarized later in this article. And there are at least seven features of data that influence the significance level of the standard Student T test of  $H_0: \rho = 0$ . The result is that great care must be taken when interpreting  $r$ . Again,  $r$  provides a useful indication of how well the least squares regression line fits a scatterplot of points, but for various reasons reviewed in this article, a collection of modern tools are needed to better detect and describe associations.

The goal in this article is to summarize the problems associated with Pearson's correlation, and least squares regression, and then mention some of the tools that might be used to address them. Many new tools for studying and characterizing associations have been developed in recent years that have the potential to greatly enhance personnel selection in particular and psychological measurement in general. It might be thought that some concerns about Pearson's correlation become irrelevant with a sufficiently large sample size but, as will be seen, there are several practical problems for which this is not always true.

## Background Information

Historically, robustness first referred to the problem of controlling the probability of a Type I error when testing a hypothesis. Today, however, it means much more. For instance, modern statisticians have derived ways of characterizing the robustness of population parameters such as the population mean  $\mu$ , the average of all individuals of interest if only they could be measured. Population parameters are said to be robust if arbitrarily small changes in a distribution cannot have an arbitrarily large effect on their value. There are three mathematical methods for assessing robustness that are at the heart of the modern theory of robustness (e.g., Huber 1981; Staudte and Sheather 1990; Wilcox 1997). By any three of these methods,  $\mu$  and the population variance  $\sigma^2$  are not robust,

Address for correspondence:  
Dept of Psychology, University  
of Southern California, USA.  
email: rwilcox@rcf.usc.edu

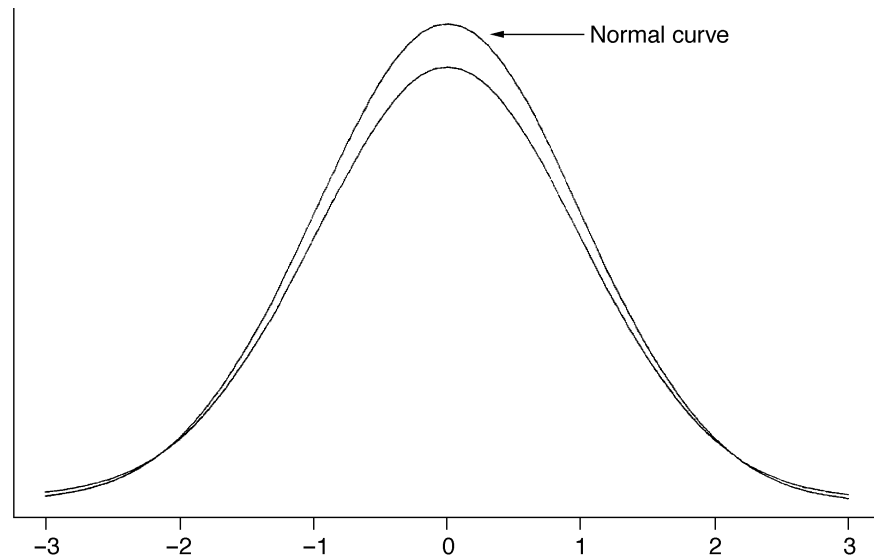


Figure 1. Plot of the standard and contaminated normal curves

and this has serious practical implications for Pearson's correlation and conventional measures of reliability, as will be seen.

Rather than go into details about mathematical issues, one of the fundamental problems is illustrated as was done in a classic paper by Tukey (1960). Figure 1 shows two symmetric probability curves. One is the standard normal, so its variance is  $\sigma^2 = 1$ . The other is not a normal curve but rather a mixed or contaminated normal. The particular mixed normal considered here is the distribution we get if we sample an observation from a standard normal with probability .9, otherwise we sample from a normal curve having standard deviation 10. As is evident, there is little visible difference between the two curves.

For some constant  $c$ , let  $P(X < c)$  be the probability that a randomly sampled observation  $X$  is less than  $c$ . The two curves in Figure 1 are similar in the sense that if we sample from the mixed normal rather than the normal curve, then  $P(X < c)$  would not be altered by more than .04 for any  $c$  we might pick. (In formal terms, their Kolmogorov distance is small.) Here is the point: Although the standard normal has a variance of 1, the variance of the mixed normal is 10.9! This illustrates an important result: *Arbitrarily small departures from normality can have an arbitrarily large impact on the population variance.* (The proof of this statement stems from a device that is similar in nature to how the mixed normal is constructed; see, for example, Staudte and Sheather, 1990.) That is, it is possible to construct a curve that is even more like the normal curve in Figure 1, yet the variance is larger than 10.9 – the variance can be made as large as we want. In modern terminology, the population variance is not robust because an

arbitrarily small change in any probability distribution can greatly affect its value.

There are many interesting implications associated with the result just described. Before turning to the main topic of this article, three of them are briefly indicated. First, from basic principles, when we sample a single observation from a normal curve, the probability that it is within one standard deviation of the mean is .68. But for the mixed normal, this probability exceeds .999. So we see that even when observations appear to follow a normal curve, the standard probabilistic interpretation of the standard deviation can be grossly inaccurate. (This has practical implications for regression analyses based on standardized scores.)

Second, it might be thought that if two probability curves have equal means and equal variances, then a graph of these curves should look fairly similar, but this is not necessarily the case even when both curves are symmetric. The left panel of Figure 2 shows two symmetric curves having equal means and variances, but they differ substantially, as is evident. The right panel shows two more curves which again have equal means and variances.

Third, when testing hypotheses using sample means or least squares regression, power (the probability of rejecting when the null hypothesis is false) can be drastically lowered from what it is when sampling from a normal curve instead (e.g., Wilcox, 1997, 1998c).

### Pearson's Correlation

Now consider Pearson's correlation. The left panel of Figure 3 shows a bivariate normal distribution with  $\rho = .8$  and the right panel is a

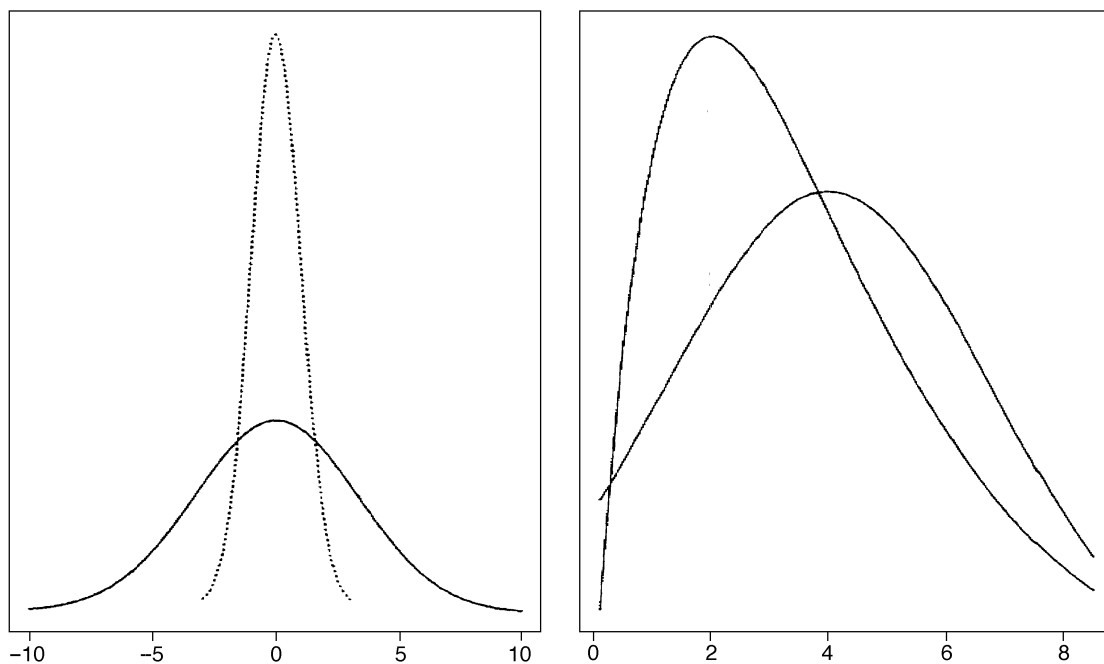


Figure 2. Distributions with equal means and variances

bivariate normal with  $\rho = .2$ . As we would expect, the two distributions are noticeably different. But now look at Figure 4. The bivariate distribution is very similar to the bivariate distribution shown in the left panel of Figure 3, but now  $\rho = .2$ , the same correlation shown in the right panel of Figure 3. In Figure 4, one of the marginal distributions is normal, but the other is a mixed normal. This illustrates that even with infinitely many observations, an arbitrarily small departure from normality might greatly affect the magnitude of  $\rho$ .

When estimating parameters based on a sample of observations, another fundamental concern is the effect of outliers. It is known that a single outlier can greatly influence the sample mean and sample variance and that the deleterious effects of outliers are much more common than once thought. The more obvious methods for dealing with this problem have been found to

be ineffective, but theoretically sound solutions have been devised (e.g., Hampel *et al.* 1986; Huber 1981; Staudte and Sheather 1990; Wilcox 1997, in press). Here the focus is on how outliers influence  $r$ .

Figure 5 shows a scatterplot of the light intensity of some stars versus their surface temperature. As is evident, the bulk of the observations appear to have a positive association, yet  $r = -.21$ . The reason is that the points in the upper left portion of Figure 5 are outliers that greatly influence  $r$ . One might argue that perhaps for  $X < 4.1$ , the association changes substantially versus  $X > 4.1$ . But there are only six points with  $X < 4.1$ , so it is difficult to know for sure. It is noted, however, that if we restrict the range of the  $X$  values by considering only  $X$  values greater than 4.1, then  $r = .65$ . As is well known, a restriction of range can lower  $r$ . Figure 5 illustrates that it can increase it as well.

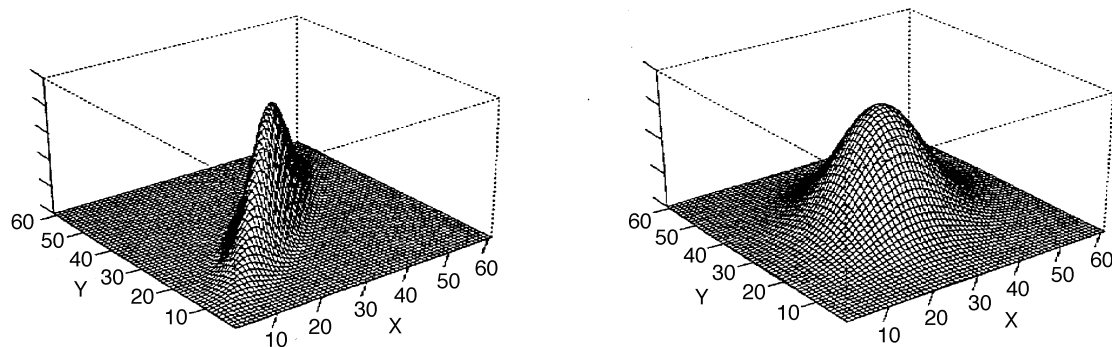


Figure 3. Two bivariate normal distributions

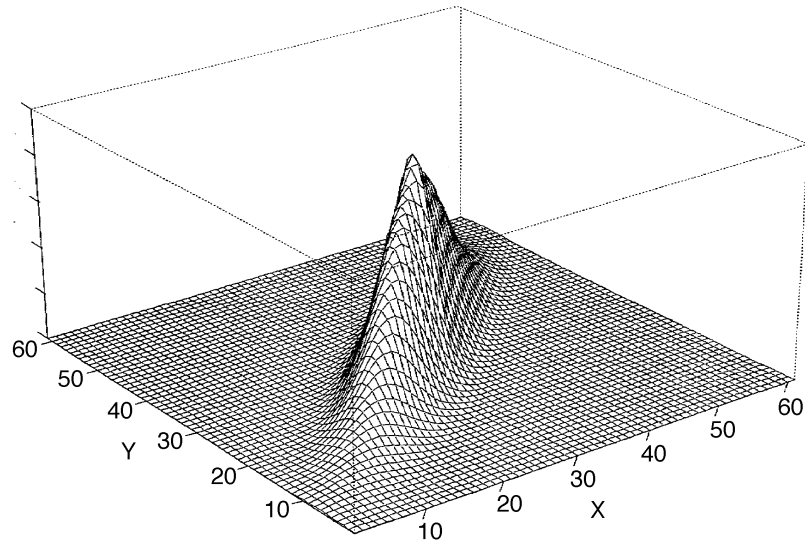


Figure 4. A bivariate distribution with correlation .2

Dealing with the problem in Figure 5 might appear to be relatively easy: Simply look at a scatterplot of points. One general problem is that detecting highly influential points that might distort the magnitude of  $r$  can be a nontrivial task. Consider, for example, Figure 6 which shows data taken from a study on predictors of reading ability. (The data are from an unpublished study by L. Doi and are reproduced in Wilcox, 1998c.) The straight nearly horizontal line is the least squares regression line having slope  $-.021$ . Moreover,  $r = -.035$  and has a significance level of .76 using the conventional Student's T test of  $H_0: \rho = 0$ .

Now look at Figure 7. Shown are the same data in Figure 6, but with two features added. The first is a so-called relplot (proposed by

Goldberg and Iglewicz 1992) which is characterized by the two ellipses. The inner ellipse contains half of the points, and points outside the outer ellipse are declared outliers. (Relplots are a bivariate generalization of a boxplot. The s-plus function `relfun`, described in Wilcox (1997) was used to create Figure 7.) The other added feature is the ragged line which is called a running interval smoother (created by the s-plus function `runmean` in Wilcox 1997). It is just one example of a collection of so-called smoothers that attempt to approximate a regression line without forcing it to have a particular shape. Hastie and Tibshirani (1990) provide an excellent introduction to smoothers. Notice that for the bulk of the points, the smoother suggests that there is a negative association. This is confirmed by any of several

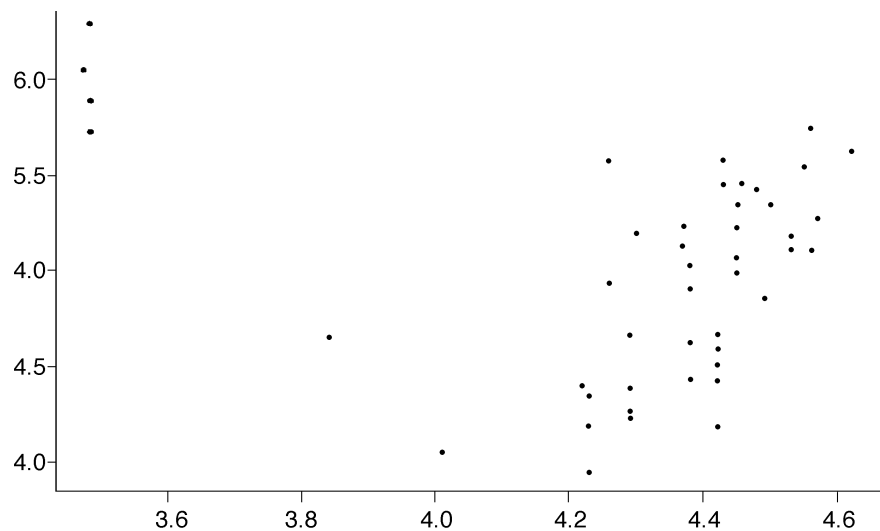


Figure 5. Star data,  $r = -0.21$

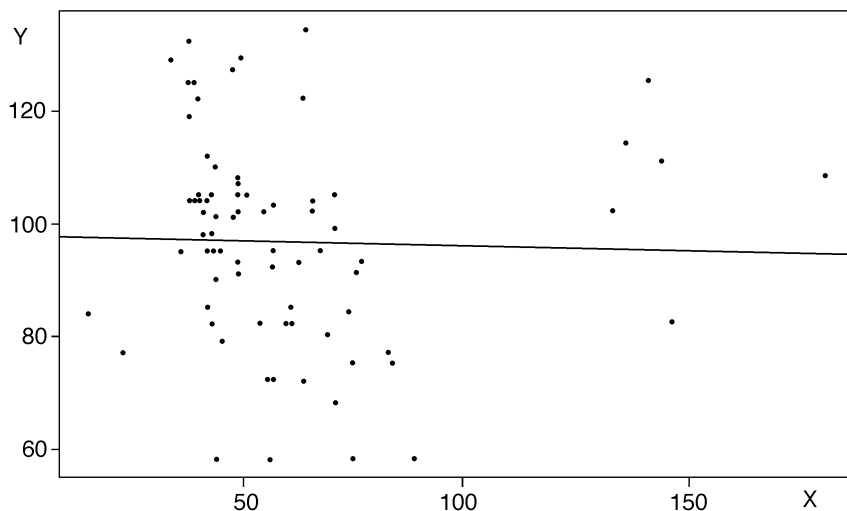


Figure 6. Data taken for a study on predicting reading ability

modern techniques that allow heteroscedasticity, portions of which are mentioned later in this article. (These methods have been found to control Type I error probabilities over a broader range of situations versus more conventional methods.)

The data in Figures 6 and 7 provide yet another illustration that a restriction of range can increase  $r^2$ . In Figure 6,  $r = -.035$ , but restricting the range of  $X$  to values less than 100,  $r$  decreases to  $-.39$ . These illustrations reflect a fundamental concern: Generally, even a single outlier among the  $X$  values can have a large impact on  $r$ . The same is true of outliers among the  $Y$  values.

Of course, curvature also affects the magnitude of  $r$ . Common strategies for dealing with curvature are to replace  $X$  with  $X^2$ , or  $\log(X)$ , or  $1/X$ . But experience with smoothers indicates that another type of curvature seems to

be more common than might be expected. Often there is an association between  $X$  and  $Y$  over some range of  $X$  values, but the association might change abruptly outside this range, and it might disappear altogether. Data from a study of diabetes in children illustrate this point. Figure 8 shows the age of children versus the logarithm of their C-peptide levels at diagnosis. (The data are from Sockett *et al.* 1987.) Because  $r = .4$  with a significance level of .008, a temptation is to conclude that C-peptide levels increase with age. But also shown in Figure 8 are the least squares regression line based on age less than seven, and the least squares line for children aged seven and older. For the first group, the hypothesis of a zero slope for the least squares regression line is rejected, but not for the other. So perhaps C-peptide levels increase with age, but it appears that this is true only up to about the age of seven.

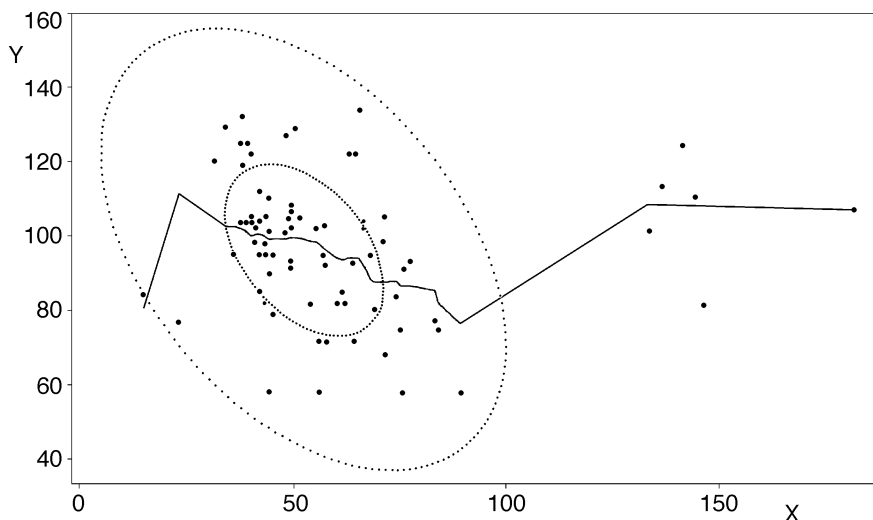


Figure 7. A replot and smooth of the data in Figure 6

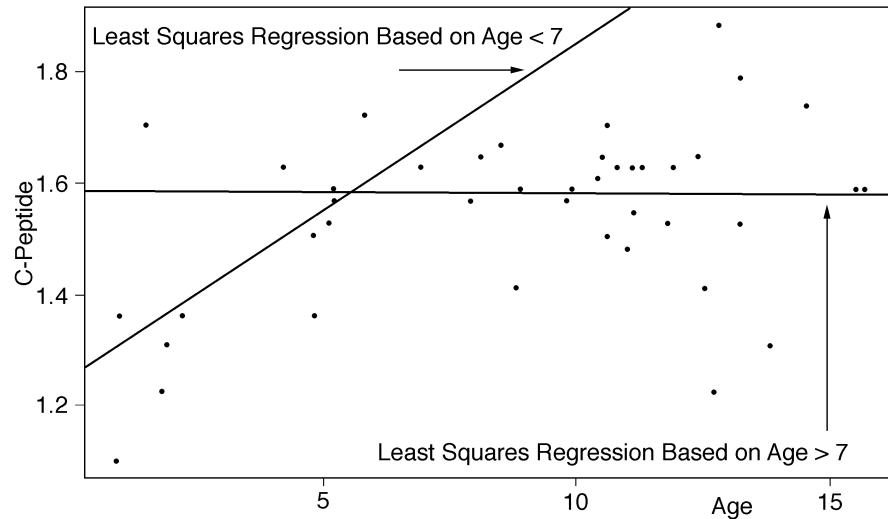


Figure 8. An illustration that associations might change abruptly

Let  $\beta_1$  be the population slope. A criticism of the conventional tests of  $H_0: \rho = 0$  and  $H_0: \beta_1 = 0$  is that they are sensitive to both heteroscedasticity and non-normality. (Heteroscedasticity refers to the conditional variance of  $Y$ , given  $X$ , changing with  $X$ . Conventional inferential methods assume homoscedasticity, meaning that the conditional variance of  $Y$  does not depend on  $X$ .) For example, it is possible to have  $\rho = 0$ , yet the probability of rejecting increases with  $n$  due to heteroscedasticity (e.g., Wilcox, in press, chapter 6). Heteroscedasticity also destroys power, even under normality, and non-normality exacerbates this problem. For these reasons, applying standard hypothesis testing methods to the data in Figure 8 might be viewed as providing a poor indication of whether the two regression lines have a slope significantly different from zero. Using more modern methods in Wilcox (1997), we again reject for the first group but not the other.

Barrett (1974) and Loh (1987) demonstrate that yet another factor influences the magnitude of  $r$ : the slope of the regression line around which points are clustered. So we see that at least six features of data are related to the magnitude of  $r$ : (1) the range of the  $X$  values; (2) reliability; (3) the magnitude of the residuals; (4) the slope of the regression line around which points are clustered; (5) curvature; and (6) outliers. Student's test of  $H_0: \rho = 0$  is sensitive to all of these features plus heteroscedasticity. Rejecting  $H_0: \rho = 0$  indicates dependence, and  $r$  indicates the extent to which  $\hat{Y} = b_1X + b_0$  improves upon  $\bar{Y}$  as a predictor of  $Y$ , where  $b_1$  and  $b_0$  are the least squares estimates of the slope and intercept. But knowing  $r$  only, a great deal of ambiguity remains about how well  $X$  predicts  $Y$  for the bulk of the points under study and the range of the  $X$  values for which a reasonably

accurate estimate of  $Y$  might be obtained. Even the sign of  $r$  might be misleading because as was demonstrated, the bulk of the points can have a positive association even when  $r$  is negative.

### Unsatisfactory Approaches to Dealing with Outliers

There are a variety of effective tools for dealing with outliers when measuring association. But to develop an appreciation for these methods, it helps to briefly describe some simple strategies that have been found to be unsatisfactory.

One way of dealing with the effects of influential outliers is to simply hope that they rarely occur. But Tukey (1960) predicted that outliers are common in applied work, and modern outlier detection methods have supported his claim. All indications are that effective methods for dealing with outliers are needed and that influential outliers occur more frequently than once was thought.

A natural strategy is to first check for outliers among the  $X$  values, and then do the same for the  $Y$  values. There are two crucial issues when considering this approach. To describe the first, consider how one might detect outliers among the  $X$  values only. Based on fundamental principles about the normal curve, a common strategy is to declare the value  $X$  an outlier if it is more than two standard deviations away from the sample mean. That is, declare the value  $X$  an outlier if:

$$|X - \bar{X}| > 2s,$$

where  $\bar{X}$  and  $s$  are the usual sample mean and standard deviation, respectively. So if we observe the values:

$$2, 3, 4, 5, 6, 7, 8, 9, 10, 50,$$

the sample mean is  $\bar{X} = 10.4$ , the standard deviation is  $s = 14.15$ , and the value 50 is declared an outlier because  $|50 - 10.4|$  exceeds  $2 \times 14.15$ . But suppose we add another outlier by changing the value 10 to 50. Then  $|\bar{X} - 50| = 1.88s$ , so 50 would not be declared an outlier, yet surely it is unusual versus the other values. If the two largest values in this last example are increased from 50 to 100, then  $|\bar{X} - 100| = 1.89s$ , and the value 100 still would not be declared an outlier. If the two largest values are increased to 1000, even 1000 would not be flagged as an outlier! This illustrates the general problem known as *masking*. Both the sample mean and standard deviation are being inflated by the outliers which in turn masks their presence. Even with very large sample sizes, masking is an issue for reasons summarized in Wilcox (in press). What is needed is an outlier detection method based on measures of location and scale that are not themselves influenced by outliers. Several methods for dealing with masking have been proposed, and probably the best-known method is the boxplot.

The second crucial issue is detecting outliers among a scatterplot of points, particularly points that inordinately influence  $r$  or the least squares regression line. A natural strategy is to simply check for outliers among the  $X$  values using a robust method that is not subject to masking, and then do the same for the  $Y$  values. But this strategy can fail to detect influential points because it does not take into account the overall structure of the cloud of points being studied. Figure 9 illustrate this issue where 20  $X$  values were generated from a standard normal distribution, and then for each  $X$ , a value for  $Y$  was generated from a normal curve having mean  $X$ . (So observations were generated from the model  $Y = X + \epsilon$ , where both  $X$  and  $\epsilon$  have a

standard normal distribution.) Then two additional points were added, both of which are located at  $(2.1, -2.4)$  and appear in the lower right corner of Figure 9.

Now, if we focus on the  $X$  values only,  $X = 2.1$  is not an outlier according to a boxplot or the method in Rousseeuw and van Zomeren (1990) when simplified to handle one variable. (Details can be found in Wilcox, in press, chapter 3.) In addition,  $Y = -2.4$  is not an outlier among all 22  $Y$  values. Despite this, Figure 9 suggests that the two points at  $(2.1, -2.4)$  are unusual, they are somewhat separated from the cloud of the other 20 values, and indeed these two points have a major impact on Pearson's correlation. If we ignore these two points and compute  $r$  with the remaining 20 points, we get  $r = .443$  and a significance level of .05. Thus, there is some indication that  $X$  and  $Y$  are dependent, which is true by construction. But if we include the two unusual values we get  $r = -0.09$  suggesting there is little or no association. Yet, there is a positive association for the bulk of the points. One could argue that the two unusual points provide evidence for a weaker association as opposed to a situation where they are not included. Surely this argument has merit, and the decrease in  $r$ , when the two outlying points are included, reflects this. However, to conclude there is no association is misleading as well. In fact,  $X = 2.1$  is the largest of the 22  $X$  values, and if we restrict the range of  $X$  to values less than 2.1, we again get  $r = .44$ , giving a strikingly different indication about the association between  $X$  and  $Y$ .

One point should be strongly stressed: It is not being suggested that outliers are uninteresting or somehow uninformative. Obviously there is interest in understanding why they arise, or learning more about the

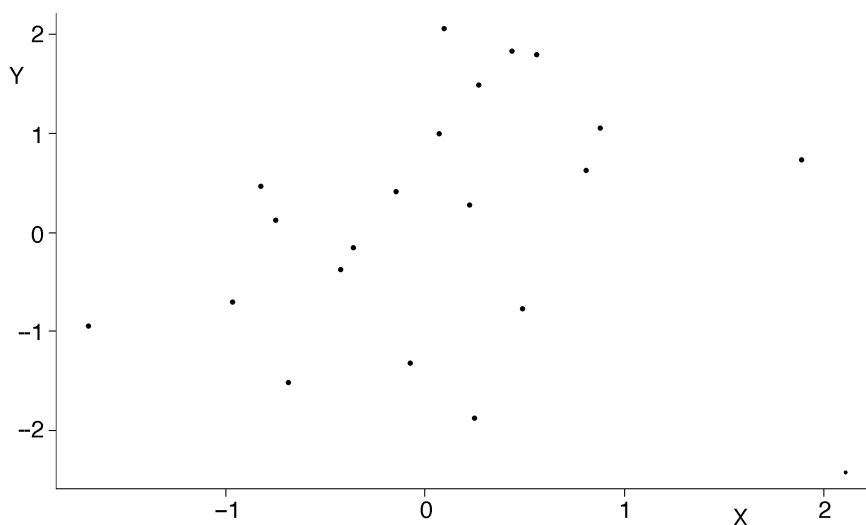


Figure 9. An illustration of how an influential outlier might be missed

nature of the individuals associated with them. They might even be the focus of attention, in which case good outlier detection methods are needed. But simultaneously, if over some range of values  $X$  provides a reasonably good estimate of  $Y$ , one may not want to miss this simply because of one or more outliers.

Another important point is that simply eliminating outliers among the  $X$  values is not always desirable or effective. One reason, as is well known, is that a restriction of range can lower  $r$ . Another general reason is that the standard error of a slope estimator can be substantially decreased. This is easily verified when using least squares, and it is true when using some robust estimators as well (e.g., Wilcox 1997). Also, simply eliminating outlying  $X$  values ignores possible problems with outlying  $Y$  values. Moreover, removing outlying  $Y$  values invalidates standard inferential techniques; special methods are required to get a valid estimate of the standard error (e.g., Wilcox 1998c, in press).

### Dealing with Outliers

There is a fairly large collection of robust correlations intended to deal with the deleterious effects of outliers that roughly fall into one of two categories. The first category is based on the general strategy of guarding against unusual  $X$  values, and then doing the same for the  $Y$  values. The second attempts to take into account the overall structure of the scatterplot of points.

Two well-known examples of the first strategy are Spearman's rho and Kendall's tau. When using Spearman's rho, converting to ranks essentially eliminates any outliers among the  $X$  values. For example, if among ten observations the largest observation is increased from 12 to one million, its rank remains unchanged, as does Spearman's rho. Of course the same is true for the  $Y$  values. In a similar manner, Kendall's tau protects against outliers among both the  $X$  and  $Y$  values. But both methods do not take into account the overall structure of a scatterplot which could lead to missing a useful association. For example, in Figure 9, ignoring the two points in the lower right, Kendall's tau and Spearman's rho are .34 and .50, respectively. But including the points, their values drop to .10 and .13. Other examples that reflect the strategy of limiting the influence of extreme  $X$  or  $Y$  values are the Winsorized correlation, the percentage bend correlation, and the biweight midcorrelation, (e.g., Wilcox 1997). The first two can be used to test the hypothesis of independence and they have an interpretation that is similar to the usual interpretation of  $r$ . The latter correlation is useful in regression when there is interest in a

robust estimate of the slope. But none of these correlations take into account the overall structure of the data. Their main advantages are that they can give a better indication, compared to  $r$ , of how the majority of points are associated (whether  $Y$  tends to increase or decrease with  $X$ ), and in some situations they provide much more power than Pearson's correlation when testing the hypothesis of a zero correlation.

Examples of robust covariance matrices that take into account the overall structure of the points are the minimum volume ellipsoid (MVE) estimator (Rousseeuw and Leroy 1987) and the minimum covariance determinant (MCD) estimator (Rousseeuw and van Driessen 1999). These covariance matrices yield robust correlations in an obvious manner. Both provide very useful tools for detecting outliers in multivariate data, and they can be used to detect influential points when fitting a straight line to data (Rousseeuw and van Zomeren 1990). Both s-plus and SAS have built-in functions for computing these robust covariance matrices. (For s-plus functions that employ these methods, see Wilcox 1997.) There are also regression methods that take into account the structure of data, a recent example of which is the deepest regression line (Rousseeuw and Hubert 1999). There are various ways these regression methods could be used to define new measures of correlation. In so far as we want to say something about whether the most central points have a positive or negative association, the sign of these robust correlations is useful. But interpreting the magnitude of these correlations is fraught with problems because there are still several features of the data that influence them. We need more than a robust analog of  $\rho$  to understand the association between  $X$  and  $Y$ .

### Robust Regression

As previously indicated, even a single outlier might grossly distort the least squares regression line, but there is another practical issue that should be mentioned: both non-normality and heteroscedasticity can cause the standard error of the least squares estimator to be substantially higher – sometimes hundreds of times higher – compared to modern robust estimators one might use (e.g., Wilcox 1997, 1998a, 1998b). Indeed, even under normality, heteroscedasticity can mean that there is a considerable practical advantage to using another estimator when testing hypotheses or computing confidence intervals.

There are many regression estimators that compete well with least squares when there is both normality and homoscedasticity, and they

can offer a substantial advantage when either of these conditions is violated. Unfortunately, it is impossible to identify the best regression estimator, the one that should routinely be used to the exclusion of all others. There is one strategy, however, that is clearly unsatisfactory: use least squares and assume all is well.

The following criteria have been found to be important in applied work. First, as argued by Huber (1993), the breakdown point of an estimator should be at least .2, where the breakdown point refers to the proportion of outliers required to destroy an estimator. The least squares estimator has a breakdown point of only  $1/n$  meaning that a single unusual point can cause the estimated slope to be arbitrarily large or small, regardless of what the other points might be. Of course, we might encounter situations where a breakdown point less than .2 is satisfactory, but to be safe, at some point an estimator with a breakdown point of at least .2 should be checked. Simultaneously, any estimator should compete reasonably well with least squares under normality and homoscedasticity, and it should offer a substantial advantage in some situations where one (or both) of these conditions is violated.

For simple regression (meaning one predictor), the following methods have been found to satisfy these criteria: the Theil-Sen estimator (Wilcox 1998a), the least trimmed squares (LTS) estimator (Rousseeuw and Leroy 1987), the least trimmed absolute (LTA) value estimator (see Hawkins and Olive 1999, for recent results), the deepest regression line (Rousseeuw and Hubert 1999), and the so-called adjusted M-estimator (Wilcox 1996, 1997). The Theil-Sen estimator has a breakdown point of .29, meaning that to give a distorted estimate of how the bulk of the points are related, more than 29% of the points must be outliers. All of these methods can be extended to multiple regression, but the breakdown point of the Theil-Sen estimator drops well below .2. Methods based on robust correlations seem to have a reasonable breakdown point with two predictors, but it also appears that the breakdown point decreases as the number of predictors increases. (A formal proof has not been found.) As for the LTS and LTA estimators, the breakdown point can be chosen by the investigator. It currently seems that in terms of efficiency (achieving a relatively small standard error), a good choice for general use is a breakdown point of .2 or .25. Moreover, it generally seems that LTA is less efficient than LTS. But in a paper to be submitted, situations are found where LTA is more efficient than LTS and where a breakdown point close to .5 is better than a breakdown point of .2 or .25. In fact, for some patterns of heteroscedasticity, a breakdown point of .5 is best even when there is

normality. The only certainty is that among modern robust estimators, no single method dominates. Although LTS can be much more accurate than least squares, and although there is weak evidence that LTS with a breakdown point of .2 or .25 is usually satisfactory, it seems wise to consider several estimators.

Confidence intervals and hypothesis testing methods are available when using modern estimators (e.g., Wilcox 1997). Currently, a method that stands out is a percentile bootstrap technique that allows heteroscedasticity. In fact, a slight modification of the method gives relatively good control over the probability of a Type I error when using least squares, even under rather extreme departures from normality and a fair degree of heteroscedasticity (Wilcox, 1996). In contrast, the conventional method can be extremely unsatisfactory. For example, when testing at the .05 level, the actual Type I error probability can exceed .5! Similar problems arise when testing  $H_0: \rho = 0$  with Student's T. In fact, under heteroscedasticity – even when there is normality – it is possible for the probability of rejection to increase with  $n$  even though  $\rho = 0$  (e.g., Wilcox, in press). The reason is that the derivation of Student's T assumes homoscedasticity, and when there is heteroscedasticity, the wrong standard error is being used. One of the important advantages of modern robust estimators is that when combined with an appropriate bootstrap method, often it is possible to detect associations missed by more conventional techniques.

A well-known method for making inferences about Pearson's correlation is to use Fisher's  $r$ -to- $z$  transformation, but it is known that this approach is not robust (e.g., Sievers 1996). In fact, Duncan and Layard (1973) describe general conditions under which the method is not even asymptotically correct, meaning that it converges to the wrong answer as the sample size increases.

### Predictive Validity

It is remarked that another approach to summarizing how well a regression line performs is its so-called prediction error, an issue that is relevant to predictive validity. Let  $(Y_i, X_i)$ ,  $i = 1, \dots, n$ , be a random sample from some bivariate distribution and suppose  $\hat{Y}$  is some estimate of  $Y$ , given  $X$ , based on these  $n$  points. An important issue in some circumstances is not how well the regression line fits the data, but how well it will perform when predicting  $Y$  with future values of  $X$ . For example, if we randomly sample a new pair of observations, say  $(X_{n+1}, Y_{n+1})$ , and if we predict  $Y_{n+1}$  with  $\hat{Y}_{n+1}$  using the regression line based on the

original  $n$  points, naturally there will be some discrepancy between  $Y_{n+1}$  and  $\hat{Y}_{n+1}$  on average. We can characterize this discrepancy with  $Q = E(|Y_{n+1} - \hat{Y}_{n+1}|)$ , the expected value of their absolute difference. Of course, we might use a squared difference instead. Effective bootstrap methods for estimating  $Q$  are available and might be considered. Currently, the so-called .632 estimator of Efron (1983) appears to perform relatively well; see also Efron and Tibshirani (1993). If, for example, attention is restricted to children less than seven years old in Figure 8, the .632 estimate of  $Q$ , when using the least squares regression, is .113, for Theil-Sen it is .098, and for LTS it is again .098. So although all three regression lines are fairly similar, there is some indication that Theil-Sen or LTS is preferable.

### Conclusion

Space limitations preclude a complete discussion of issues and modern developments, but hopefully this article will help stimulate consideration of modern methods in personnel selection and related areas. A general suggestion is that a variety of robust and graphical tools be used to analyze data. Smoothers, for example, can be an invaluable tool for detecting situations where a restriction of range might reveal an association that is not otherwise detected. Relying exclusively on Pearson's correlation is a method that is outdated and a practice that needs to be reconsidered. A practical issue is deciding whether least squares regression should be used at all. The answer seems to depend on what we hope to accomplish. In terms of predicting  $Y$ , given  $X$ , if the least squares regression line is highly similar to some robust regression line, it would seem that least squares might be satisfactory. However, in terms of hypothesis testing, despite any similarity there might be, a robust method can have substantially more power. An obvious way of avoiding modern methods is to use diagnostic tools in an attempt to justify standard techniques. But all indications are that diagnostic tools do not always have enough power to detect practical problems. Currently, the only way to know whether it makes a difference if modern tools are used is to try both. It is recommended, however, that when making inferences about regression parameters, the modern methods in Wilcox (1997) be used exclusively. (Currently, the only known method for dealing effectively with heteroscedasticity is to use a particular bootstrap technique that is motivated by theoretical results in Wu 1986.)

### References

- Barrett, J.P. (1974) The coefficient of determination: some limitations. *Annals of Statistics*, **28**, 19–20.
- Duncan, G.T. and Layard, M.W. (1973) A monte-carlo study of asymptotically robust tests for correlation. *Biometrika*, **60**, 551–8.
- Efron, B. (1983) Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of the American Statistical Association*, **78**, 316–31.
- Efron, B. and Tibshirani, R.J. (1993) *An Introduction to the Bootstrap*. New York: Chapman and Hall.
- Goldberg, K.M. and Iglewicz, B. (1992) Bivariate extensions of the boxplot. *Technometrics*, **34**, 307–20.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J. and Stahel, W.A. (1986) *Robust Statistics*. New York: Wiley.
- Hastie, T.J. and Tibshirani, R.J. (1990) *Generalized Additive Models*. New York: Chapman and Hall.
- Hawkins, D.M. and Olive, D. (1999) Applications and algorithm for least trimmed sum of absolute deviations regression. *Computational Statistics and Data Analysis*, **32**, 119–34.
- Huber, P.J. (1981) *Robust Statistics*. New York: Wiley.
- Huber, P.J. (1993) Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti and W. Stahel (eds) *New Directions in Statistical Data Analysis and Robustness*, Boston: Birkhäuser Verlag, 139–46.
- Loh, W.-Y. (1987) Does the correlation coefficient really measure the degree of clustering around a line? *Journal of Educational Statistics*, **12**, 235–9.
- Rousseeuw, P.J. and Hubert, M. (1999) Regression depth. *Journal of the American Statistical Association*, **94**, 388–402.
- Rousseeuw, P.J. and Leroy, A.M. (1987) *Robust Regression and Outlier Detection*. New York: Wiley.
- Rousseeuw, P.J. and van Driessen, K. (1999) A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, **41**, 212–23.
- Rousseeuw, P.J. and van Zomeren, B.C. (1990) Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 633–9.
- Sievers, W. (1996) Standard and bootstrap confidence interval for the correlation coefficient. *Journal of Mathematical and Statistical Psychology*, **49**, 381–96.
- Socket, E.B., Daneman, D., Carlson, C. and Ehrlich, R.M. (1987) Factors affecting and patterns of residual insulin secretion during the first year of type 1 (insulin dependent) diabetes mellitus in children. *Diabetes*, **30**, 453–9.
- Staudte, R.G. and Sheather, S.J. (1990) *Robust Estimation and Testing*. New York: Wiley.
- Tukey, J.W. (1960) A survey of sampling from contaminated normal distributions. In I. Olkin *et al.* (eds.) *Contributions to Probability and Statistics*. Stanford, CA: Stanford University Press.
- Wilcox, R.R. (1996) Confidence intervals for the slope of a regression line when the error term has non-constant variance. *Computational Statistics and Data Analysis*, **22**, 89–98.
- Wilcox, R.R. (1997) *Introduction to Robust Estimation and Hypothesis Testing*. San Diego, CA: Academic Press.

- 
- Wilcox, R.R. (1998a) A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal*, **40**, 261–8.
- Wilcox, R.R. (1998b) Simulation results on extensions of the Theil-Sen regression estimator. *Communications in Statistics: Simulation and Computation*, **27**, 1117–26.
- Wilcox, R.R. (1998c) The goals and strategies of robust methods. *British Journal of Mathematical and Statistical Psychology*, **51**, 1–39.
- Wilcox, R.R. (2001) *Fundamentals of Modern Statistical Methods*. New York: Springer-Verlag.
- Wu, C.F.J. (1986) Jackknife, bootstrap, and other resampling methods in regression. *Annals of Statistics*, **14**, 1261–1295.