

Inferences About a Robust Heteroscedastic Measure of Effect Size When There Are Two Covariates

Rand R. Wilcox

Correspondence: Dept of Psychology, University of Southern California, USA

Received: January 23, 2024 Accepted: April 24, 2024 Online Published: May 29, 2024

doi:10.5539/ijsp.v13n2p46

URL: <https://doi.org/10.5539/ijsp.v13n2p46>

Abstract

Recently, a method was proposed for making inferences about a robust, heteroscedastic measure of effect size when there is a covariate. The method is readily extended to two covariates, but nothing is known about how well it controls the Type I error probability. This note reports results indicating why, when dealing with two covariates, the Type I error probability drops well below the nominal level when the sample sizes are small. A modification of the method is suggested that performs well in simulations. The method is used to compare two groups of participants who are categorized as depressed or not depressed. The dependent variable is a measure of meaningful activities. The two covariates are a measure of stress and a measure of perceived health.

Keywords: linear model, quantile regression estimator, bootstrap, robust effect size

1. Introduction

A well-known goal is characterizing how two independent groups compare based on a measure of effect size that is a function of some measure of location as well as some measure of dispersion. Let μ_j and σ_j denote the mean and variance, respectively, associated with the j th group ($j = 1, 2$). Certainly, one of the best-known versions is

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}, \quad (1)$$

where it is assumed that $\sigma = \sigma_1 = \sigma_2$. That is, homoscedasticity is assumed.

A fundamental concern with Δ is that it is not robust. Imagine, for example, that when dealing with normal distributions, $|\Delta|$ is considered to be small, medium or large if $|\Delta| = 0.2, 0.5$ and 0.8 , respectively, as suggested by Cohen (1988). Algina et al. (2005) demonstrate that a very small shift from a normal distribution to a heavy-tailed distribution can substantially lower $|\Delta|$. This result is related to the fact that an arbitrarily small departure from a normal distribution can inflate the variance substantially (e.g., Staudte & Sheather, 1990). A related concern is the impact of outliers on the sample mean and variance. A few outliers can mask a large effect size among the bulk of the participants. Algina et al. dealt with this concern by replacing the mean and variance with a 20% trimmed mean and Winsorized variance.

Another issue is heteroscedasticity. That is, how might one deal with situations where $\sigma_1 \neq \sigma_2$? Let n_j denote the sample size for the j th group. Let $N = n_1 + n_2$ and $\ell = n_1/N$. Kulinskaya et al. (2008) deal with this issue by replacing Δ with

$$\delta = \frac{\mu_1 - \mu_2}{s}, \quad (2)$$

where

$$s^2 = \frac{(1 - \ell)\sigma_1^2 + \ell\sigma_2^2}{\ell(1 - \ell)}.$$

A simple robust analog of δ is to follow the approach used by Algina et al. (2005). That is, replace the mean and variance with a 20% trimmed mean and Winsorized variance. The computational details are not described because they are not relevant here for reasons made obvious in the next section. The only point is that (2) explains the motivation for the robust, heteroscedastic measure of effect size described in section 2.

Generally, a fundamental goal is making inferences about some measure of effect size when there is a covariate. Wilcox (2022a) proposed such a method when using a variation of (2). The method does not assume parallel regression lines as done by the classic ANCOVA method and it allows heteroscedasticity. However, the results were limited to a single covariate. The method is easily extended to two covariates but there are no simulation results on how well it performs. One goal in this note is to fill this gap. As will be seen, having two covariates impacts the ability of the method to control the probability of a Type I error; the actual level can be substantially less than the nominal level for reasons to be described. The main goal is to suggest a modification of the method that deals with this issue.

The paper is organized as follows. Section 2 describes how the method in Wilcox (2022a) is readily extended to two covariates. Section 3 reports simulation results indicating that this obvious extension is unsatisfactory when the sample sizes are relatively small. This is followed by a suggested modification that addresses this concern. Section 4 applies the method using data dealing with the physical and emotional wellbeing of older adults.

2. Description of the Method

For convenience, momentarily consider a single group and assume that the q quantile of Y , given that $X_1 = x_1$ and $X_2 = x_2$, is given by

$$\beta_{2,q}x_2 + \beta_{1,q}x_1 + \beta_{0,q} \quad (3)$$

Given a random sample (Y_i, X_{1i}, X_{2i}) ($i = 1, \dots, n$), the parameters $\beta_{0,q}$, $\beta_{1,q}$ and $\beta_{2,q}$ can be estimated using the method derived by Koenker and Bassett (1978). Letting r_i denote the residuals, this is done by minimizing the function

$$\sum \psi_q(r_i), \quad (4)$$

where

$$\psi_q(u) = u(q - I_{u < 0}), \quad (5)$$

and the indicator function $I_{u < 0} = 1$ if $u < 0$, otherwise $I_{u < 0} = 0$.

Next, let

$$V_q(x_1, x_2) = b_{2,q}x_2 + b_{1,q}x_1 + b_{0,q} \quad (6)$$

denote the estimate of the q quantile of Y given that $(X_1, X_2) = (x_1, x_2)$. Then a robust estimate of a conditional measure of variation is

$$U(x_1, x_2) = \frac{V_{0.75}(x_1, x_2) - V_{0.25}(x_1, x_2)}{z_{0.75} - z_{0.25}}, \quad (7)$$

where $z_{0.75}$ and $z_{0.25}$ are the 0.75 and 0.25 quantiles, respectively, of a standard normal distribution. That is, $U(x_1, x_2)$ is an estimate of the conditional interquartile range that is rescaled to estimate the conditional population standard deviation when the conditional distribution of Y is normal. As is evident, the conditional median of Y is estimated with

$$M(x_1, x_2) = b_{0.5,2}x_2 + b_{0.5,1}x_1 + b_{0.5,0}. \quad (8)$$

Now, for two independent groups, let $M_j(x_1, x_2)$ and $U_j(x_1, x_2)$ denote $M(x_1, x_2)$ and $U(x_1, x_2)$, respectively, for the j th group. Then a robust heteroscedastic analog of the measure of effect size given by (2) is estimated with

$$\hat{\eta}(x_1, x_2) = \frac{M_1(x_1, x_2) - M_2(x_1, x_2)}{\hat{\varphi}(x_1, x_2)}, \quad (9)$$

where

$$\hat{\varphi}(x_1, x_2)^2 = \frac{(1 - \ell)U_1^2(x_1, x_2) + \ell U_2^2(x_1, x_2)}{\ell(1 - \ell)}$$

and again $\ell = n_1/(n_1 + n_2)$.

Let $\eta(x_1, x_2)$ denote the population analog of $\hat{\eta}(x_1, x_2)$ and consider the goal of testing

$$H_0 : \eta(x_1, x_2) = 0 \quad (10)$$

and computing a confidence interval for $\eta(x_1, x_2)$. Following Wilcox (2022a), first compute a bootstrap estimate of the standard error of $\hat{\eta}(x_1, x_2)$. For notational convenience, consider a random sample from a single group:

$$\begin{pmatrix} X_{11}, X_{12}, Y_1 \\ \vdots \\ X_{n1}, X_{n2}, Y_n \end{pmatrix}. \quad (11)$$

The method begins by randomly sampling with replacement n rows from this matrix yielding $(Y_1^*, X_{11}^*, X_{21}^*), \dots, (Y_n^*, X_{1n}^*, X_{2n}^*)$. Here, a bootstrap sample is generated from each group and $\eta(x_1, x_2)$ is estimated based on these bootstrap samples yielding $\hat{\eta}^*(x_1, x_2)$. Repeat this process B times yielding $\hat{\eta}_1^*(x_1, x_2), \dots, \hat{\eta}_B^*(x_1, x_2)$. An estimate of the standard error of $\hat{\eta}(x_1, x_2)$ is

$$S^2(x_1, x_2) = \frac{1}{B-1} \sum (\hat{\eta}_b^*(x_1, x_2) - \bar{\eta}^*(x_1, x_2))^2, \quad (12)$$

where $\bar{\eta}^*(x_1, x_2) = \sum \hat{\eta}_b^*(x_1, x_2)/B$ (e.g., Efron & Tibshirani, 1993).

An issue is the choice for B , the number of bootstrap samples. A natural reaction is that B should be reasonably large, say 1000 or larger. Here, however, $B = 1000$ resulted in extremely high execution times when running simulations. Results in Efron (1987) suggest that $B = 100$ suffices and the results reported here are based on this choice in order to avoid high execution times. Some simulations were run with $B = 1000$. This impacted the Type I error rate by a few units in the third decimal place. Nevertheless, in practice $B = 1000$ might be preferred because it improves the stability of the results if a different seed for a random number generator is used.

Inferences about $\eta(x_1, x_2)$ are made assuming that

$$W = \frac{\hat{\eta}(x_1, x_2) - \eta(x_1, x_2)}{S(x_1, x_2)} \quad (13)$$

has a standard normal distribution. From basic principles, for the common goal of testing (10), reject at the α level if $|\hat{\eta}(x_1, x_2)|/S(x_1, x_2) \geq z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution. A $1 - \alpha$ confidence interval for $\eta(x_1, x_2)$ is

$$\hat{\eta}(x_1, x_2) \pm z_{1-\alpha/2} S(x_1, x_2). \quad (14)$$

2.1 Choosing Covariate Points

In practice, there might be substantive reasons to focus on particular covariate points. The approach used here is to choose points based on selected quantiles of the marginal distributions. Yet another possibility is use the so-called projection depth of the covariate points, which can be computed using methods described in Wilcox (2022b, section 6.2.5). This is basically the method used by Wilcox (2023) when dealing with an analog of the Wilcoxon–Mann–Whitney method. Briefly, this approach quantifies how deeply each point is nested within the cloud of covariate points yielding say $\mathcal{D}_1, \dots, \mathcal{D}_N$. This approach does not assume that the distribution is elliptically contoured in contrast to robust analogs of Mahalanobis distance. The higher the depth, the closer is the point to the center of the data cloud. (The R function `pdepth` in the R package WRS performs the calculations.) The strategy is to use a sequence of $N < n$ points based on the depth measures so that points near the center of the cloud, as well as points near the edge of the cloud, are included.

3. Simulation Results

Simulations were used to assess the Type I error probability when testing (10). Here, three covariate points are used. The first is the point corresponding the lower quartiles of the two marginal distributions associated with the first group. The second is the median of the marginal distributions followed by the upper quartiles. These three points are denoted by Pt1, Pt2 and Pt3, respectively.

The first set of simulations were based on three types of distributions: standard normal, a lognormal distribution shifted to have a median of zero, and a symmetric heavy-tailed distribution. The lognormal distribution is commonly used in simulations (e.g., Wilcox, 2022) and it is further motivated by results reported by Cain et al. (2017) who gathered 1,567 estimates of skewness and kurtosis reported in various published papers. The skewness and kurtosis of a lognormal distribution are 6.185 and 113.9, respectively. These values are larger than 99% of the estimates reported by Cain et al. (2017). This suggests that if a method performs reasonably well for this seemingly large departure from a normal distribution, this offers some assurance that it will work well in practice. Here, a heavy-tailed symmetric distribution is included to get some sense of the impact of moving from a symmetric distribution to a skewed distribution.

More precisely, data were generated from a bivariate distribution where the marginal distributions have a g-and-h distribution. This was done via the R function `ghmul` in the R package WRS. The parameters g and h determine the first four moments (Hoaglin, 1985). The choice $(g, h) = (0, 0)$ corresponds to a standard normal distribution, $(g, h) = (1, 0)$ corresponds to a lognormal distribution shifted to have a median of zero, while $(g, h) = (0, 0.2)$ is a symmetric distribution about zero having kurtosis approximately equal to 24.8. For recent results on the properties of g-and-h distributions, see Astivia and Edward (2022).

The actual Type I error probability when testing (10) at the .05 level was estimated based on 3000 replications. Initial estimates indicated that the actual level drops well below 0.025 when using $n_1 = n_2 = 20$ and even 30. Some results are reported here for sample sizes $(n_1, n_2) = (50, 50)$ and $(50, 100)$ in order to illustrate that even now, the actual level can be rather unsatisfactory. Two choices for ρ were used, namely 0 and 0.5, where ρ is Pearson's correlation. As can be seen, for $(n_1, n_2) = (50, 50)$ the lowest estimate is 0.013. Bradley (1978) suggested that when testing at the .05 level, the actual level should be at least 0.025 and no larger than 0.075. When $n_2 = 100$, the lowest estimates is 0.018. Note that the estimates are fairly consistent as a function of the distributions used as well as the choice for ρ .

Some simulations were run with $(n_1, n_2) = (100, 100)$ and $(400, 400)$ as a partial check on how the method performs

with large sample sizes. For $g = h = 0, \rho = 0$ and $(n_1, n_2) = (100, 100)$ the estimates were 0.038, 0.029 and 0.024. For $(n_1, n_2) = (400, 400)$ the estimates were 0.044, 0.039 and 0.038.

Table 1. Estimated Type I error probabilities

| ρ | g | h | n_1 | n_2 | Pt1 | Pt2 | Pt3 |
|--------|-----|-----|-------|-------|-------|-------|-------|
| 0.0 | 0.0 | 0.0 | 50 | 50 | 0.025 | 0.019 | 0.018 |
| 0.0 | 1.0 | 0.0 | 50 | 50 | 0.028 | 0.019 | 0.017 |
| 0.0 | 0.0 | 0.2 | 50 | 50 | 0.025 | 0.017 | 0.020 |
| 0.5 | 0.0 | 0.0 | 50 | 50 | 0.025 | 0.014 | 0.013 |
| 0.5 | 1.0 | 0.0 | 50 | 50 | 0.028 | 0.021 | 0.020 |
| 0.5 | 0.0 | 0.2 | 50 | 50 | 0.027 | 0.014 | 0.015 |
| 0.0 | 0.0 | 0.0 | 50 | 100 | 0.029 | 0.024 | 0.024 |
| 0.0 | 1.0 | 0.0 | 50 | 100 | 0.035 | 0.023 | 0.024 |
| 0.0 | 0.2 | 0.2 | 50 | 100 | 0.027 | 0.019 | 0.018 |
| 0.5 | 0.0 | 0.0 | 50 | 100 | 0.030 | 0.024 | 0.023 |
| 0.5 | 1.0 | 0.0 | 50 | 100 | 0.035 | 0.023 | 0.023 |
| 0.5 | 0.0 | 0.2 | 50 | 100 | 0.028 | 0.027 | 0.022 |

An issue is why the estimates tend to be well below the nominal level when the sample sizes are small. One possible explanation is that $\hat{\eta}(x_1, x_2)$ is not approximately normal. However, a Q-Q plot of $\hat{\eta}(x_1, x_2)$ based on 10000 estimates of $\eta(x_1, x_2)$ when $g = h = \rho = 0$ and $n_1 = n_2 = 50$ suggests that to a reasonable degree, the distribution of $\hat{\eta}(x_1, x_2)$ is normal. However, there is a serious problem with the bootstrap estimate of the standard error. The 10000 estimates used to create the Q-Q plot provide an estimate of the standard error of $\hat{\eta}(x_1, x_2)$. This value was compared to bootstrap estimates of the standard error revealing that the estimate is biased, the estimates tend to be higher than the actual standard error, which explains why the estimated Type I error probabilities tend to be less than the nominal level. Increasing the number of bootstrap samples to $B = 1000$ did not correct this problem. A percentile bootstrap method does not require an estimate of the standard error, but it did not improve the control over the Type I error. Using the jackknife method derived by Quenouille (1949) to estimate the bias based on the bootstrap values offered no advantage.

A commonly used strategy is to find a method that performs reasonably well under normality and then investigate the impact of using this approach when sampling from a non-normal distribution. The approach used here is to momentarily assume normality, $\rho = 0$, and determine a constant $c(n)$ so that $S_a = c(n)S$ is approximately unbiased when $x_1 = x_2 = 0$. This was done for $n = 20(10)50(25)200, 250, 300$ using the approach in the previous paragraph. The resulting estimates of $c(n)$ are shown in Table 2.

Table 2. Estimates of $c(n)$

| n | $c(n)$ |
|-----|--------|
| 20 | 0.067 |
| 22 | 0.713 |
| 25 | 0.756 |
| 30 | 0.805 |
| 31 | 0.808 |
| 34 | 0.825 |
| 40 | 0.845 |
| 50 | 0.870 |
| 75 | 0.897 |
| 100 | 0.923 |
| 125 | 0.936 |
| 150 | 0.940 |
| 175 | 0.944 |
| 200 | 0.945 |
| 250 | 0.954 |
| 300 | 0.952 |

Figure 1 shows a plot of $c(n)$ versus $1/n$. One could fit a linear model with the goal of determining $c(n)$ based on $1/n$.

However, it was found that the non-parametric smoother derived by Cleveland (1979) performed a bit better. For sample sizes not included in Table 1, values for $c(n)$, when $20 \leq n \leq 300$, can be computed via the R function `lplot.pred` in the R package WRS.

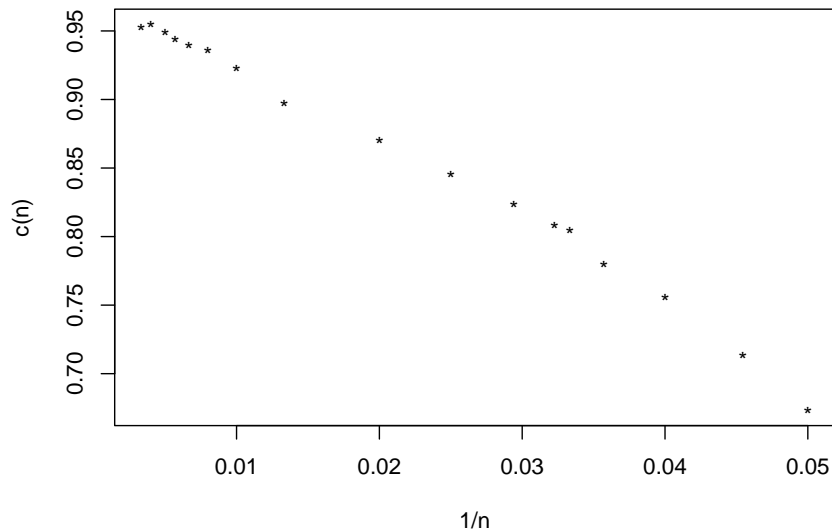


Figure 1. A plot of $c(n)$ versus $1/n$

Table 3 shows the simulation results based on the modified estimate of the standard error. As can be seen, control over the Type I error probability is substantially better compared to the results in Table 1. When the smallest sample size is 20, the lowest estimate was .024 in one situation only. All indications are that the actual level is closer to the nominal level when using $S_a = c(n)S$ to estimate the standard error rather than S . In addition, all indications are that Bradley's criterion is satisfied.

Table 3. Estimated Type I error probabilities using adjusted estimates of the standard error

| ρ | g | h | n_1 | n_2 | Pt1 | Pt2 | Pt3 |
|--------|-----|-----|-------|-------|-------|-------|-------|
| 0.0 | 0.0 | 0.0 | 20 | 20 | 0.030 | 0.066 | 0.041 |
| 0.0 | 1.0 | 0.0 | 20 | 20 | 0.038 | 0.060 | 0.025 |
| 0.0 | 0.0 | 0.2 | 20 | 20 | 0.036 | 0.067 | 0.041 |
| 0.5 | 0.0 | 0.0 | 20 | 20 | 0.036 | 0.070 | 0.048 |
| 0.5 | 1.0 | 0.0 | 20 | 20 | 0.048 | 0.060 | 0.026 |
| 0.5 | 0.0 | 0.2 | 20 | 20 | 0.033 | 0.068 | 0.046 |
| 0.0 | 0.0 | 0.0 | 20 | 50 | 0.032 | 0.052 | 0.043 |
| 0.0 | 1.0 | 0.0 | 20 | 50 | 0.032 | 0.042 | 0.024 |
| 0.0 | 0.0 | 0.2 | 20 | 50 | 0.027 | 0.046 | 0.040 |
| 0.5 | 0.0 | 0.0 | 20 | 50 | 0.035 | 0.055 | 0.041 |
| 0.5 | 1.0 | 0.0 | 20 | 50 | 0.040 | 0.049 | 0.026 |
| 0.5 | 0.0 | 0.2 | 20 | 50 | 0.031 | 0.052 | 0.043 |
| 0.0 | 0.0 | 0.0 | 50 | 50 | 0.043 | 0.054 | 0.045 |
| 0.0 | 1.0 | 0.0 | 50 | 50 | 0.045 | 0.059 | 0.036 |
| 0.0 | 0.0 | 0.2 | 50 | 50 | 0.044 | 0.058 | 0.046 |
| 0.5 | 0.0 | 0.0 | 50 | 50 | 0.049 | 0.055 | 0.047 |
| 0.5 | 1.0 | 0.0 | 50 | 50 | 0.051 | 0.060 | 0.039 |
| 0.5 | 0.0 | 0.2 | 50 | 50 | 0.045 | 0.060 | 0.043 |
| 0.0 | 0.0 | 0.0 | 50 | 100 | 0.046 | 0.050 | 0.043 |
| 0.0 | 1.0 | 0.0 | 50 | 100 | 0.045 | 0.056 | 0.036 |
| 0.0 | 0.0 | 0.2 | 50 | 100 | 0.044 | 0.051 | 0.040 |
| 0.5 | 0.0 | 0.0 | 50 | 100 | 0.049 | 0.052 | 0.043 |
| 0.5 | 1.0 | 0.0 | 50 | 100 | 0.049 | 0.052 | 0.035 |
| 0.5 | 0.0 | 0.2 | 50 | 100 | 0.047 | 0.054 | 0.044 |

4. An Illustration

The method is illustrated with data stemming from the Well Elderly study (Clark et al. 2012) that dealt with the emotional and physical wellbeing of older adults. Two groups are formed based on a measure of depressive symptoms. A score greater than 15 is generally interpreted as an indication of mild depression or worse. The first group consists of participants having scores less than or equal to 15 and the second group consists of participants having scores greater than 15. The dependent variable of interest is a measure of meaningful activities (MAPA) and the two covariates are measures of stress and a measure of perceived health (PH). The sample sizes are 297 and 161. The groups were compared by first combining the covariate points and computing the projection depth of all $297+161=458$ points. Next, covariate depths were put in ascending order and then the groups were compared based on twenty covariate points that are evenly spaced between the deepest point and the least deep point. Figure 2 shows the results. Points marked by * indicate the covariate points that were used. Points marked by o are the covariate points where (10) is rejected at the .05 level using unadjusted p-values. That is, the p-values were not adjusted to control the familywise error (FWE) rate (the probability of one or more Type I errors).

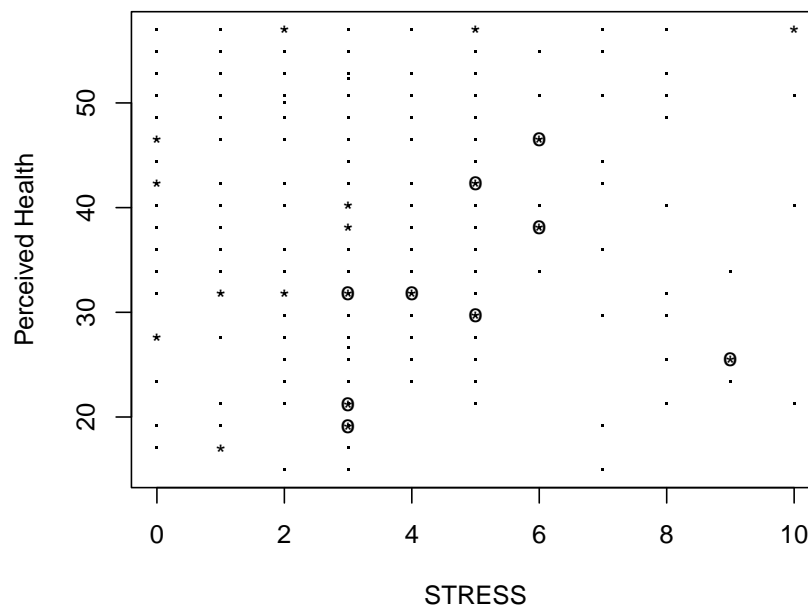


Figure 2. Results based on the Well Elderly data

Table 4 summarizes the results for the covariate points where the p-value is less than or equal to .05. The column headed p.adjusted are p-values adjusted to control the FWE rate using a method derived by Hochberg (1988). As can be seen, five points remain significant when the p-values are adjusted. The median stress value is 4 and the proportion of stress values less than or equal to 3 is .45. The median perceived health is 38. Roughly, when stress is near 38 or higher, and perceived health is near or below 38, MAPA scores tend to be significantly higher for the first group.

Perhaps more importantly, the results provide a perspective beyond merely saying that a significant result was obtained. As previously noted, a common convention is to view $|\Delta| = 0.2, 0.5$ and 0.8 as being a small medium and large effect size, respectively. These values correspond to $|\eta| = 0.1, 0.25$ and 0.4 . Most of the estimates in Table 4 are small by this standard and in general the lower ends of the confidence intervals would not rule out a very small effect size. The largest estimate is 0.434 , which is relatively large and corresponds to the stress measure equal to 9 and the perceived health measure equal to 25.47 . That is, a relatively large effect is indicated when stress is relatively large and perceived health is relatively low. The lower end of the confidence interval would not rule out a medium effect size, but the upper end would not rule out an extremely large effect size.

Table 4. Estimated effect sizes that are significant (unadjusted p-values less than .05) using the Well Elderly data

| Est. | Test.Stat | ci.low | ci.up | p-value | p.adjusted |
|-------|-----------|--------|-------|---------|------------|
| 0.483 | 30.802 | 0.234 | 0.731 | 0.001 | 0.003 |
| 0.173 | 10.976 | 0.001 | 0.345 | 0.048 | 0.577 |
| 0.163 | 20.014 | 0.004 | 0.322 | 0.043 | 0.571 |
| 0.179 | 20.809 | 0.054 | 0.304 | 0.005 | 0.074 |
| 0.226 | 40.083 | 0.118 | 0.335 | 0.000 | 0.001 |
| 0.216 | 30.715 | 0.102 | 0.331 | 0.000 | 0.004 |
| 0.113 | 20.140 | 0.010 | 0.216 | 0.033 | 0.451 |
| 0.150 | 30.383 | 0.063 | 0.237 | 0.001 | 0.012 |
| 0.157 | 30.123 | 0.059 | 0.256 | 0.002 | 0.028 |

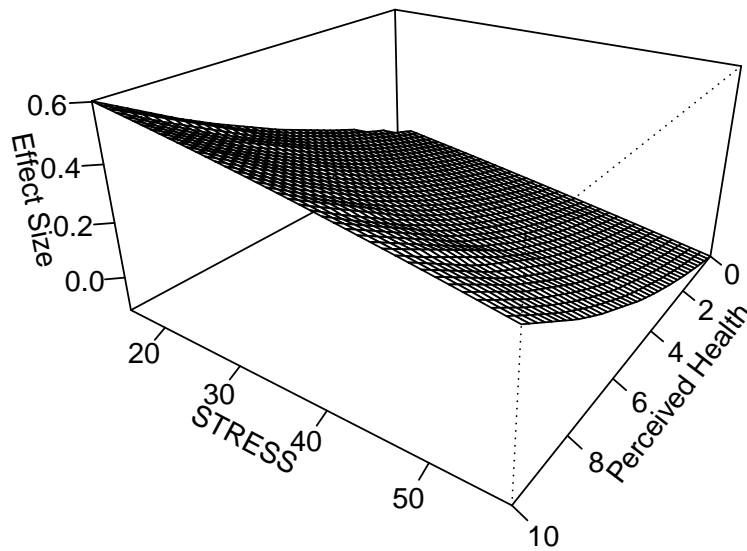


Figure 3. Results based on the Well Elderly data

Notice that the KMS measure of effect size, $\eta(x_1, x_2)$, can be estimated for all of the covariate values and plotted using a nonparametric regression method. Figure 3 shows such a plot based on the smoother derived by Cleveland and Devlin (1988) again using the Well Elderly data. The plot suggests that the effect size increases as perceived health increases and stress decreases and that the magnitude of the effect size ranges between relatively low values and values that are quite high.

5. Concluding Remarks

There are, of course, other approaches to comparing groups that are not based on a measure of effect size that is a function of both a measure of location and scale. Certainly the best-known approach is to use just a measure of location. It is not being argued that the approach used here should be preferred over other approaches that might be used. As noted by Steegan et al. (2016), multiple methods can be required to get a nuanced understanding of how groups compare. The suggestion is that the method studied here can help achieve this goal.

All indications are that the proposed method, based on the adjusted estimate of the standard error, performs reasonably well in terms of Type I error probabilities. Moreover, the adjustments aimed at achieving a more unbiased bootstrap estimate of the standard error, which was derived assuming normality, appears to perform well when dealing with non-normal distributions. The extent this remains the case when there are more than two covariates remains to be determined.

It is fairly evident that no single regression estimator is always optimal. The quantile regression is no exception. For example, an issue raised by Mosteller and Tukey (1977, p. 366) is that when dealing with the 0.5 quantile, a relatively large weight is being given to observations with the smallest residuals.

Finally, the R function `ancovap2.KMS` computes the KMS measure of effect size for two covariates. And the R function `ancovap2.KMSci` computes confidence intervals. The argument `nboot` can be used to specify B , the number of bootstrap samples. The plot shown in Figure 3 was created with the R function `ancovap2.KMS.plot`. These functions are stored in the file `Rallfun-v43`, which can be downloaded from <https://osf.io/xhe8u/> as well as <https://zenodo.org/records/10420647>.

Reference

- Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10, 317-328.
- Astivia, O., & Edward, K. (2022). Theoretical considerations when simulating data from the g-and-h family of distributions. *British Journal of Mathematical and Statistical Psychology*, 75, 699-727. <https://doi.org/10.1111/bmsp.12274>
- Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152. <https://doi.org/10.1111/j.2044-8317.1978.tb00581.x>
- Cain, M., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavioral Research*, 49, 1716-1735. <https://doi.org/10.3758/s13428-016-0814-1>
- Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh M., ... Azen, S. P. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health*, 66, 782-790.
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74, 829-836. <https://doi.org/10.2307/2286407>
- Cleveland, W. S., & Devlin, S. J., (1988) Locally-weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83, 596-610. <https://doi.org/10.2307/2289282>
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. (2nd Ed). New York: Academic Press.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82, 171-185. <https://doi.org/10.1080/01621459.1987.10478410>
- Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall. <https://doi.org/10.1007/978-1-4899-4541-9>
- Hoaglin, D. C. (1985). Summarizing shape numerically: the g-and-h distribution. In: Hoaglin, D., Mosteller, F., Tukey, J. (Eds.), *Exploring Data Tables Trends and Shapes*. New York: Wiley, pp. 461-515.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75, 800-802. <https://doi.org/10.1093/biomet/75.4.800>
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrika*, 46, 33-50. <https://doi.org/10.2307/1913643>
- Kulinskaya, E., Morgenthaler, S., & Staudte, R. (2008). *Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence*. New York: Wiley. <https://doi.org/10.1002/9780470985533>
- Mosteller, F., & Tukey, J. W. (1977). *Data Analysis and Regression*. Addison-Wesley, Reading, MA.
- Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, B*, 11, 18-84. <https://doi.org/10.1111/j.2517-6161.1949.tb00023.x>
- Staudte, R. G., & Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley. <https://doi.org/10.1002/9781118165485>
- Steege, S., Tuerlinck, F., Gelman, A., & Vanpaemel, W. (2016). Increasing Transparency Through a Multiverse Analysis. *Perspectives on Psychological Science*, 11, 702-712. <https://doi.org/10.1177/1745691616658637>

- Wilcox, R. (2022a). ANCOVA: An approach based on a robust heteroscedastic measure of effect size. *Sankhya: The Indian Journal of Statistics B*, 84, 831–845. <https://doi.org/10.1007/s13571-022-00291-4>
- Wilcox, R. (2022b). Introduction to Robust Estimation and Hypothesis Testing. 5th Edition. San Diego, CA: Academic Press <https://doi.org/10.1016/B978-0-12-820098-8.00007-5>
- Wilcox, R. (2023). Some results on estimating a Wilcoxon–Mann–Whitney measure of effect size when there are two covariates. *Sankhya, B*. <https://doi.org/10.21203/rs.3.rs-2870213/v1>

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).