MEDIAN OVERLAP, MISCLASSIFICATION AND R-SQ

Abstract: The median overlap is an extremely useful descriptive statistic for comparing two univariate distributions involving a continuous variable (e.g., heights of males and females). The median overlap is the percentage of subjects in one group having a value that goes past the median value for subjects in the other group (e.g., percentage of females whose heights exceed the median height for males). When the variable of interest is normally distributed with equal standard deviations in both groups, *the median overlap is closely related to the chance that the comparative relationship between a random subject from each group will be misclassified.* (If the median overlap is 16% for heights of males and females, the chance of a male being shorter than a female is 24%.) Students should be exposed to the chance of misclassification so as to improve their understanding of the discriminatory power of a particular binary variable. The median overlap determines the value of R-squared in an OLS regression. The rate of misclassification is shown to be modeled as 43% - .79*R-squared. The certainty in making a correct comparison is modeled as 57% + .78 R-squared. An R-squared of about 50% gives one about 90% certainty in making a comparative prediction.

I. INTRODUCTION:

Often our goal is to make predictions. To improve the quality of our predictions, we use the values of related variables. This is commonly done involving discrete classification variables (tables) and continuous variables (regression).

Consider a group of subjects having a quantitative property (e.g., height). Suppose these subjects can be divided into two groups by a binary variable (e.g., sex) having two values (e.g., male and female). Suppose the two distributions have medians that differ by a certain amount (X). Given this, we can identify the proportion of the one population that has values of the continuous variable that overlap past the median of the other distribution (e.g., the percentage of females who are taller than the median male). This percentage is called the median overlap (Bell Curve, p. 49). For illustration, suppose that both distributions were normal with equal standard deviations (s) and a separation distance, X. In this case the median overlap would be CDF(-X/s). So if X is 3" and the standard deviation for height in both sexes is also 3", then the median overlap should be about 16%. This is shown in the following picture.
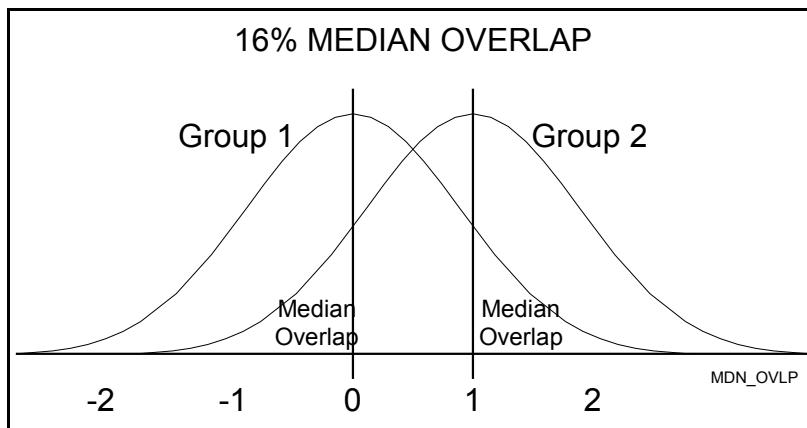

Figure 1: Median Overlap

There are two ways of expressing the median overlap in terms of probabilities.
1. The median overlap is also the chance that a random member of one group will be above (below) the median of the other group.
2. If the variable of interest is normally distributed with equal standard deviations, *the median overlap is always less than the chance of misclassification using group membership to predict the relationship between random members from each group.*
The truth of the first expression is a simple consequence of the nature of the median overlap and the equality of opportunity under a random choice. The truth of the second expression requires proof.

To understand what is being asserted, consider the following cases:

Sex and height. Assume that 16% of all females are taller than the median height of the males (a 16% median overlap). In this case, there is a 16% chance that a random male will be shorter than the average female. If heights are normally distributed, there is at least a 16% chance that a random male will be shorter than a random female.

Handedness and IQ. Assume that left-handers are slightly smarter than right-handers. Specifically, assume that 40% of all left handers have an IQ that is below the median IQ of the right handers (a 40% median overlap). In this case there is a 60% chance that the IQ of a random left hander will be higher than the IQ of a random right hander.   If these IQs are normally distributed with equal standard deviations, then there is at least a 60% chance that the IQ of a random left-hander will be greater than the IQ of a random right-hander.

II. PROOF: If two groups (males and females) have a common property (e.g., height) which is normally distributed, and if subjects are picked randomly from both groups, then the resulting distribution will be a bivariate normal. The correlation coefficient between the random selections from both groups will be zero. If the means for both groups are identical, then the joint center would be located on the 45° line. On one side of this line are those subjects which are properly classified (male is taller than female); on the other side of this line are those subjects that are misclassified (male is shorter than female). In such a case, the proportion of the bivariate normal involving the misclassification is 50% -- half of the bivariate distribution.

Now suppose the medians for the two groups are unequal. The joint center will be to one side (or the other) of this 45° line. Those subjects who are misclassified based on their sex are located on the side of the 45° line opposite the side containing the joint center.
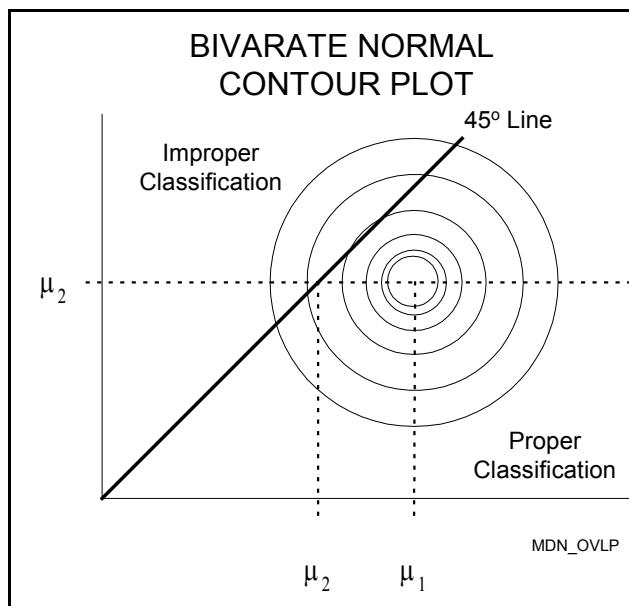
Figure 1 Median Overlap

Since the heights of males and females selected randomly are uncorrelated (rho = 0), it follows that a slice in any direction will yield a normal distribution. Consider all the slices perpendicular to the 45° line. Each perpendicular slice is cut off by the 45° line at the same value of Z_p. Thus, the proportion of the volume on the wrong side of the 45° line (equal heights) is simply the p-value associated with that value of Z_p. The value of Z_p is .707 Z where Z = X/s and where .707 is 1/sqrt(2). Thus, the percentage of misclassifications is always greater than the median overlap.

The next figure illustrates the relation between the misclassification rate and the median overlap.

The maximum difference is 8.3 percentage points; this occurs when the median overlap is 12%. Thus, median overlap ≤ misclassification rate ≤ (median overlap + 8 percentage points).

If the two groups have equal numbers then the overall group median is at Z/2. If one calculates [CDF(Z/2) + CDF(Z)]/2 = [(Joint Median Overlap) + (Opposite group median overlap)]/2, this is within 1 percentage points of the misclassification rate for all values of Z for Z ≤ 2. The misclassification rate is approximately equal to the average of two median overlap rates.

Misclassification rate versus median overlap

### Maximum of 8 percentage points difference at 12%.
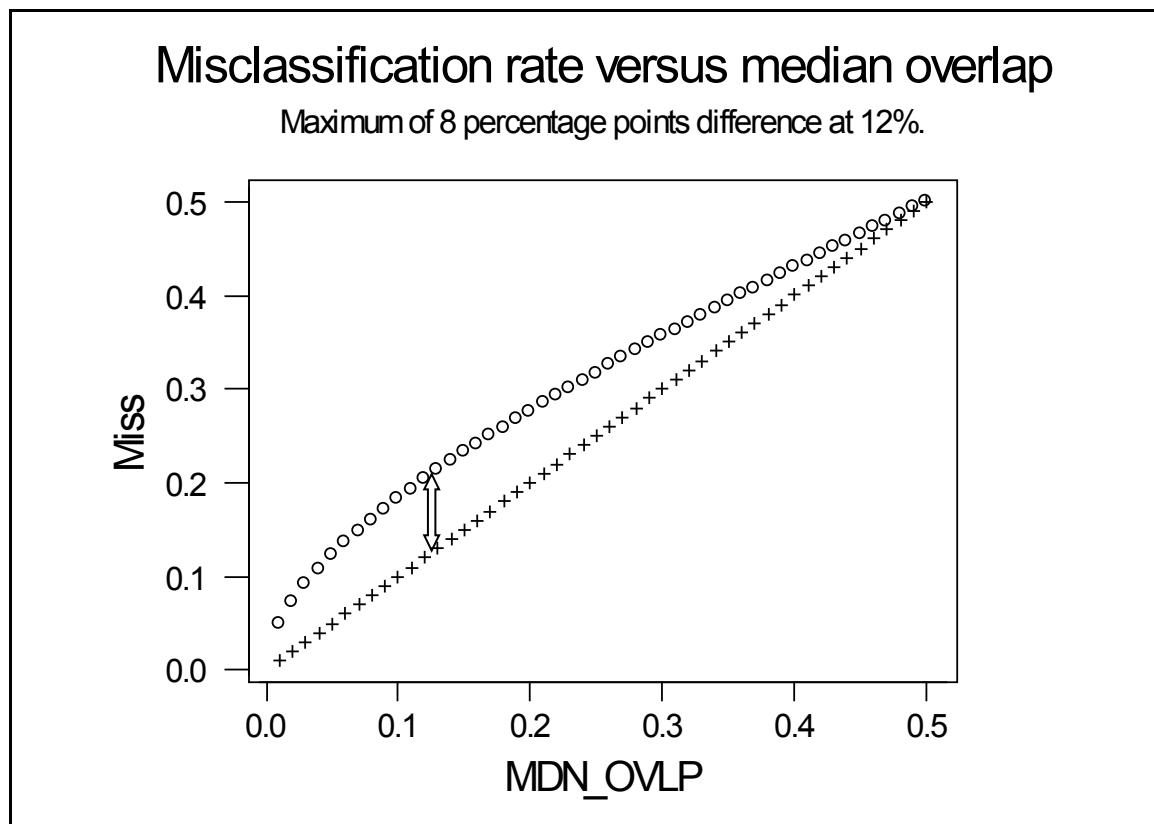
Figure 2 Misclassification vs. Median Overlap

As the median overlap decreases, the rate of misclassification decreases. These two variables are positively associated and closely associated.

III. EXAMPLES:
According to the Bell Curve (p. 49), there is a 7% median overlap in IQ between graduates of high school (who don't complete college) and graduates of college (Z=1.48). Suppose one selects a random college graduate and a random high-school graduate who didn't complete college. Assuming these IQs are normally distributed with equal standard deviations, there is a 15% chance of misclassification since $Z_p = 1.05$. Thus there is an 85% chance that the IQ of a random college graduate exceeds that of a random high school graduate.

According to the Bell Curve (p. 49), there is a 21% median overlap in IQ between those college graduates who have obtained a Ph.D., M.D., or LL.B., and those college graduates who have not (Z= .81). Assuming these IQs are normally distributed with equal standard deviations, there is a 28% chance of misclassification since $Z_p = 0.57$. Thus there is a 72% chance that the IQ of a random student receiving the advanced degree is higher than the IQ of a random student who completed college but did not receive the advanced degree.

IV. ERROR REDUCTION:

These probabilities of correct forecasts seem quite high. Remember that even if there is no predictive power for a given binary variable, one will still be right 50% of the time due strictly to chance. An alternate way to measure improvement is to calculate how much of the original uncertainty (50%) was removed for various Misclassification Rates (MR). The reduction in error is 100% (50% - MR) / MR.
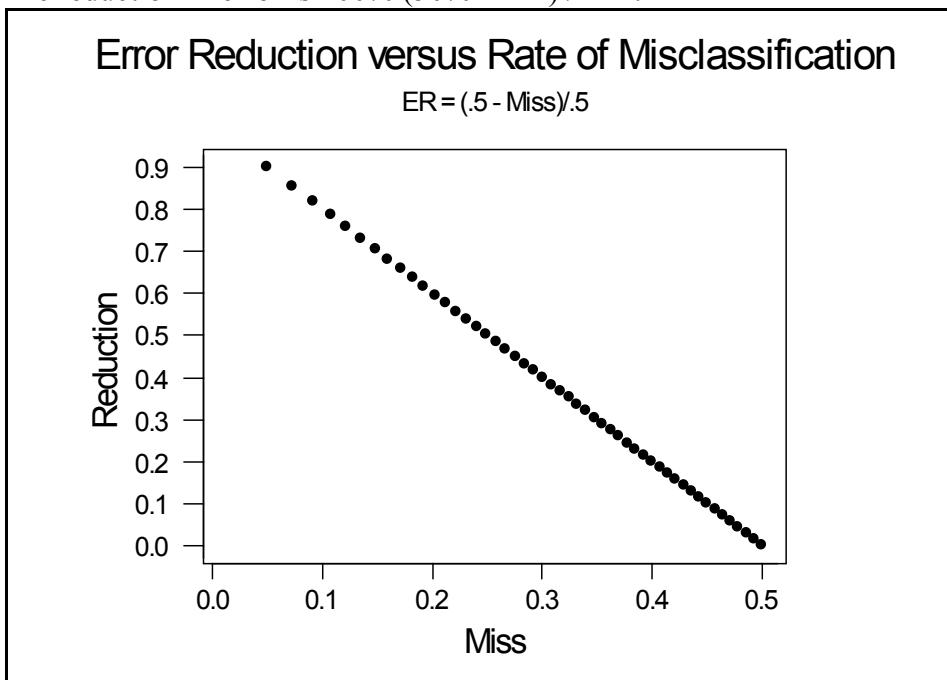
Figure 3 Error Reduction vs. Misclassification Rate

The median overlap is closely related to the rate of misclassification. Thus, as the median overlap decreases, the rate of reduction in misclassification error increases.
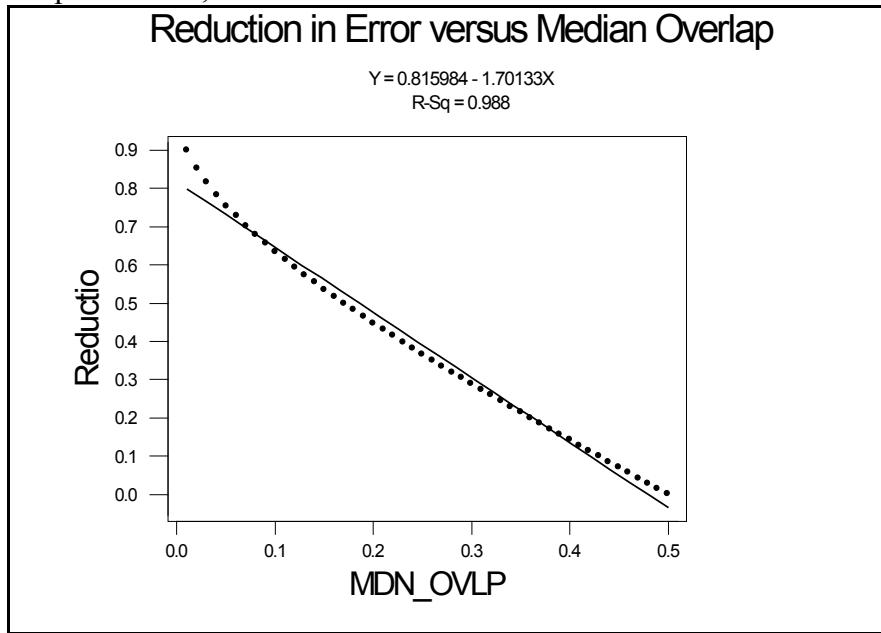
Figure 4 Error Reduction vs. Median Overlap

V. R-SQUARED:

In an ordinary least-squares regression, Rsquared is the percentage of variance eliminated by using the model. Let Z measure the separation distance between the medians of the two groups measured in standard deviations. When the variances of the two groups are equal and when the numbers in the two groups are equal, it is shown in the appendix that:

$$R^2 = (Z^2/4)/[(Z^2/4 + 1)] = 1/[(1 + 1/(Z^2/4)] \qquad \text{Eq 14.}$$

$$Z^2 = (\mu_2 - \mu_1)^2 / s^2.$$

An OLS regression using Median Overlap to predict Rsquared, gives approximately

Rsquared = 41% - Median Overlap
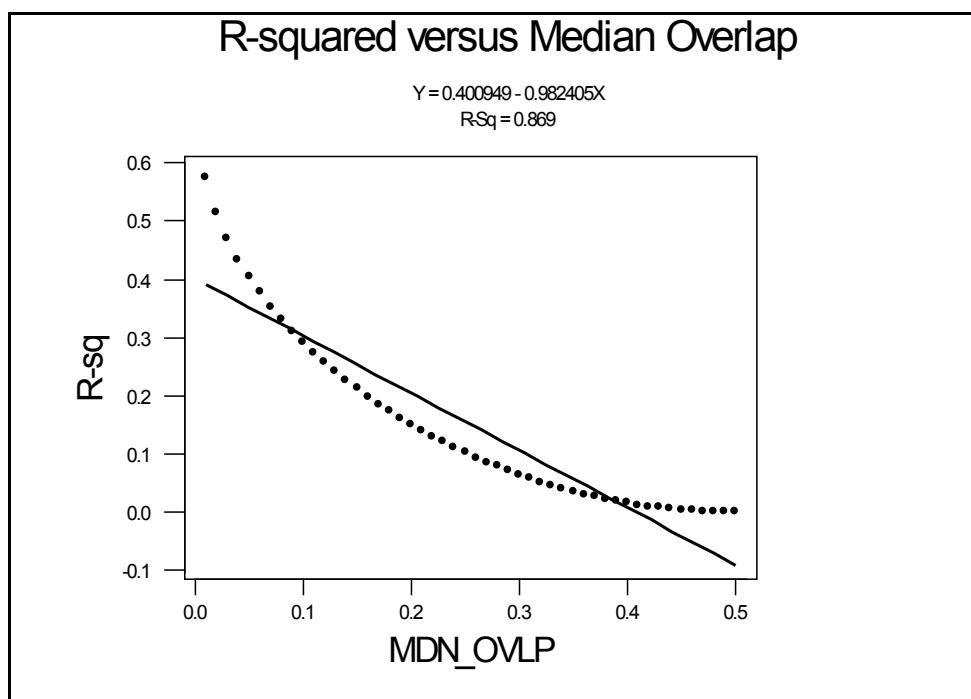
Figure 5 R-sq vs. Median Overlap

An OLS regression using Rsquared to predict Median Overlap, gives approximately

Median Overlap = 39% - .89 Rsquared

The proportion misclassified can be estimated using OLS regression on Rsquared:

Proportion Misclassified = 0.427 - 0.785 Rsquared.
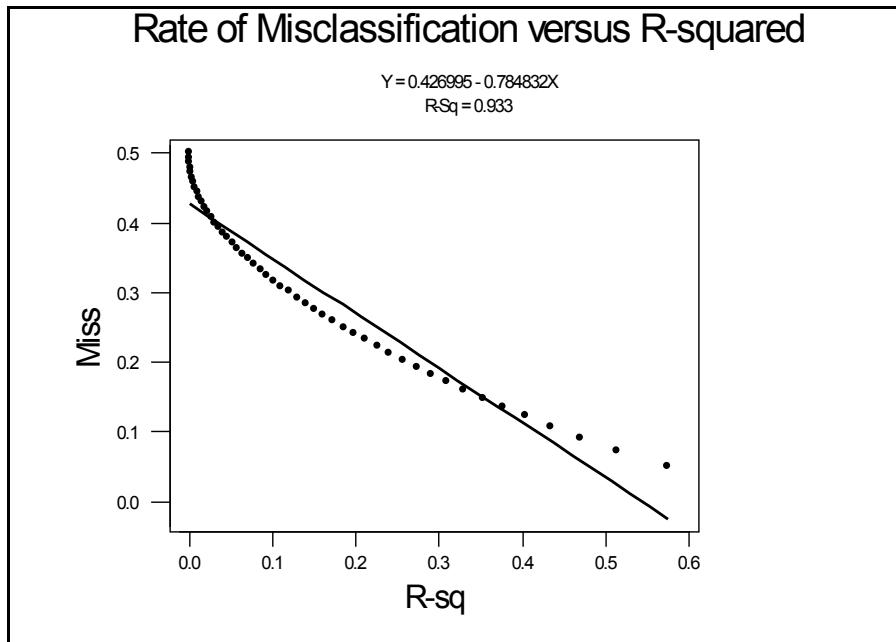
Figure 6 Misclassification Rate vs. R-sq

We can also model the certainty of our comparison.
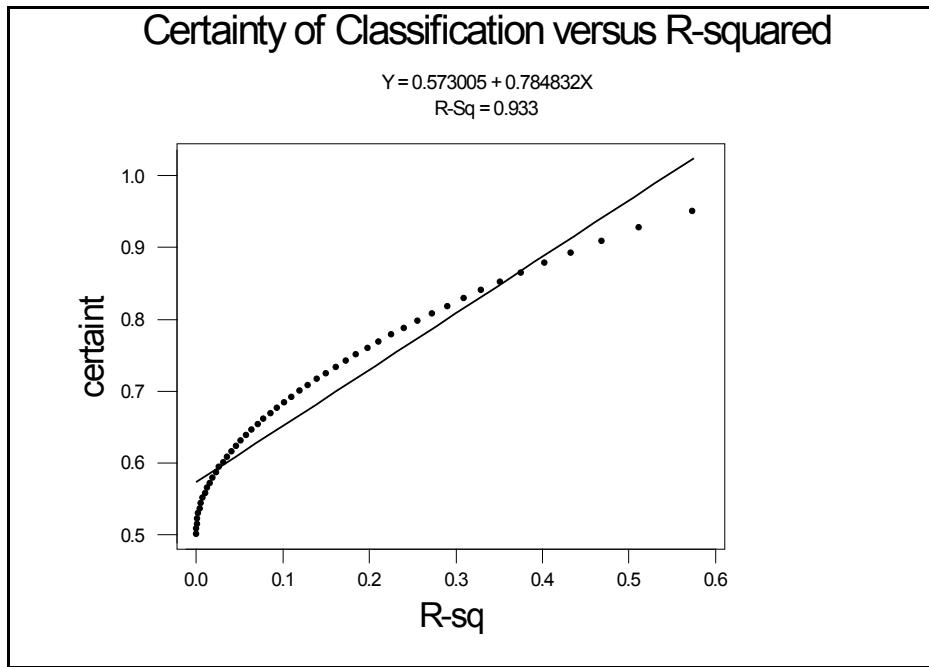
Figure 7 Certainty of Classification vs. R-squared

An R-squared of 0.5 takes one to a certainty of about 90% in making a relative classification. Thus, it doesn't take a very high R-squared to make accurate predictions involving a simple comparison.

VI. ERROR REDUCTION AND R_SQUARED

We can compare Rsquared with the rate of error reduction. This is shown in the following figure.
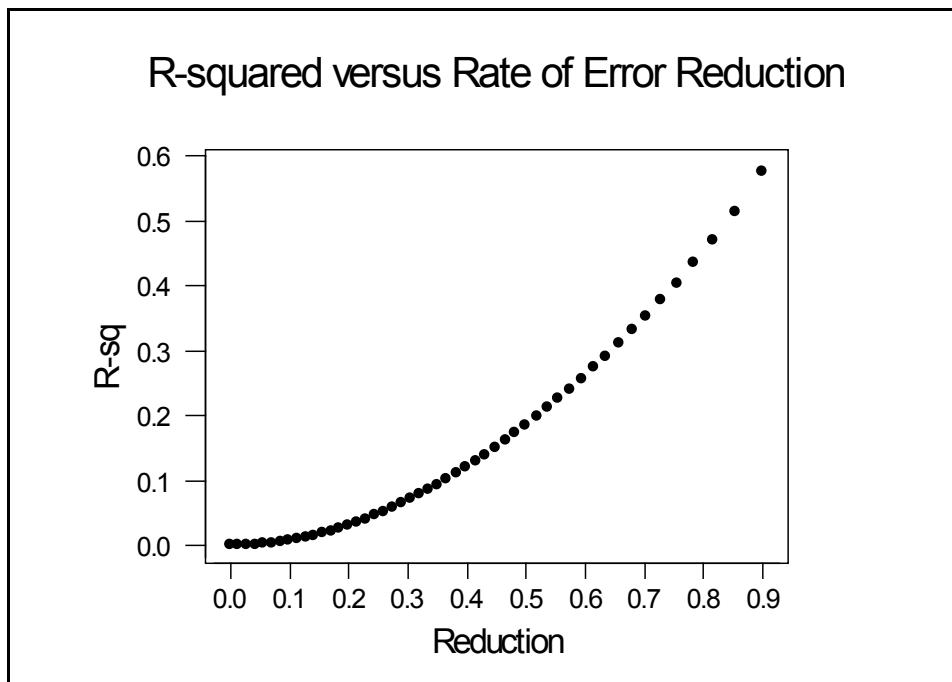
Figure 8 R-squared vs. Error Reduction Rate

VII. SUMMARY

We are justified in using the value of a binary variable (male vs. female) to make an inferential comparison of a related quantitative variable (height) when the median overlap is small. A median overlap of 12% (5%) gives an 80% (90%) rate of correct predictions. A median overlap of 1% gives a 95% rate of correct predictions.

It doesn't take a very high value of R-squared to have a 90% certainty in making a simple ordinal comparison (height of male > height of female). One reason is that by chance alone, one has a 50% chance of making such a comparison correctly. So an R-squared of zero gives one a 50% chance of a correct comparison.

To obtain a 90% chance of making a correct comparison, one needs a median overlap of no more than 12% or an R-squared of at least 37%.

APPENDIX: RELATION BETWEEN STANDARD DEVIATIONS
Thanks to Gerald Kaminski:

Consider two distributions, $X(\mu_1, \sigma_1)$ and $X(\mu_2, \sigma_2)$ having n_1 and n_2 values respectively. If these two distributions are mixed together, then the combined distribution has a center μ_0 and a standard deviation σ_0.

$$\mu_0, = (n_1 \mu_1 + n_2 \mu_2)/(n_1+n_2) \qquad\qquad\qquad\qquad\qquad \text{Eq 1.}$$

$$\sigma_0^2 = \sum (x - \mu_0)^2 /(n_1+n_2) \qquad\qquad\qquad\qquad\qquad \text{Eq 2}$$

Now divide this into two parts – one for each of the two distributions:

$$\sigma_0^2 = \sum [(x - \mu_1) + (\mu_1- \mu_0)]^2 /(n_1+n_2) + \sum [(x - \mu_2) + (\mu_2- \mu_0)]^2 /(n_1+n_2) \qquad \text{Eq 3.}$$

where each part includes only the data points from that group. Expanding, we get

$$\sigma_0^2 = \sum [(x - \mu_1)^2 + 2(x - \mu_1) (\mu_1- \mu_0) + (\mu_1- \mu_0)^2] /(n_1+n_2) + \qquad \text{Eq 4}$$
$$\sum [(x - \mu_2)^2 + 2(x - \mu_2) (\mu_2- \mu_0) + (\mu_2- \mu_0)^2] /(n_1+n_2)$$

Note that the cross-product sums to zero.

$$s_1^2 = (x - \mu_1)^2/n_1 \quad \text{and} \qquad s_2^2 = (x - \mu_2)^2/n_2 \quad \text{for the respective groups.} \qquad \text{Eq 5.}$$

$$\sigma_0^2 = \{n_1 [s_1^2 + (\mu_1- \mu_0)^2] /(n_1+n_2)\} + \{n_2 [s_1^2 + (\mu_2- \mu_0)^2] /(n_1+n_2)\} \qquad \text{Eq 6.}$$

If $s_1 = s_2 = s$, and $n_1 = n_2$, then

$$\mu_0 = (\mu_1 + \mu_2)/2 \text{ so that } \mu_0 - \mu_1 = \mu_2 - \mu_0. \text{ Thus, } \mu_2 - \mu_1 = 2 (\mu_2 - \mu_0) \qquad \text{Eq 7.}$$

$$\sigma_0^2 \quad = \quad [2s^2 + (\mu_1- \mu_0)^2 + (\mu_2- \mu_0)^2] /2 \qquad\qquad \text{Eq. 8}$$

$$\sigma_0^2 \quad = \quad [2s^2 + (\mu_2- \mu_1)^2 /2] /2 \qquad\qquad\qquad \text{Eq. 9.}$$

$$\boxed{\sigma_0^2 \quad = \quad s^2 + [(\mu_2- \mu_1)/2]^2} \qquad\qquad\qquad \text{Eq 10.}$$

$$\text{Let } Z = |\mu_2 - \mu_1| / s. \qquad \text{Thus, } s^2 = (\mu_2 - \mu_1)^2 / Z^2 \qquad\qquad \text{Eq 11.}$$

$$\sigma_0^2 \quad = \quad (\mu_2- \mu_1)^2[1/4 + 1/Z^2] = s^2Z^2 [1/4 + 1/Z^2] = s^2(Z^2/4 + 1) \qquad \text{Eq 12.}$$

===

Now consider R^2:

$$R^2 = (\sigma_0^2 - s^2)/ \sigma_0^2 = [s^2(Z^2/4 + 1) - s^2]/ [s^2(Z^2/4 + 1)] \qquad\qquad \text{Eq 13.}$$

This reduces to

$$\boxed{R^2 = (Z^2/4)/ [(Z^2/4 + 1)] = 1/ [(1 + 1/(Z^2/4)]} \qquad\qquad \text{Eq 14.}$$

For small Z, $R^2 = Z^2/4$. For large Z, $R^2 = 1 - 1/(Z^2/8)$.   Eq 15.

For s = 1 and $\mu_2 - \mu_1 = 1$, Z = 1 and $R^2 = 20\%$.   Here the median overlap is 16%.

For s = 1 and $\mu_2 - \mu_1 = 2$, Z = 2 and $R^2 = 50\%$.   Here the median overlap is 2.5%.

For s = 1 and $\mu_2 - \mu_1 = 3$, Z = 3 and $R^2 = 69\%$. Here the median overlap is 0.5%.

```
APPENDIX:

MTB > Stack c2 c3 c4;
SUBC > subscripts c5.

MTB > desc c2 c3 c4
```
Descriptive Statistics

```
Variable          N      Mean    Median   Tr Mean    StDev   SE Mean
X(0,1)           99    0.0000    0.0000    0.0000    0.9652    0.0970
X(1,1)           99    1.0000    1.0000    1.0000    0.9652    0.0970
Stack           198    0.5000    0.5000    0.5000    1.0854    0.0771

Variable        Min       Max        Q1        Q3
X(0,1)       -2.3263    2.3263   -0.6745    0.6745
X(1,1)       -1.3263    3.3263    0.3255    1.6745
Stack        -2.3263    3.3263   -0.2598    1.2598

MTB > dotplot c2 c3 c4;
SUBC> same.
```
Character Dotplot
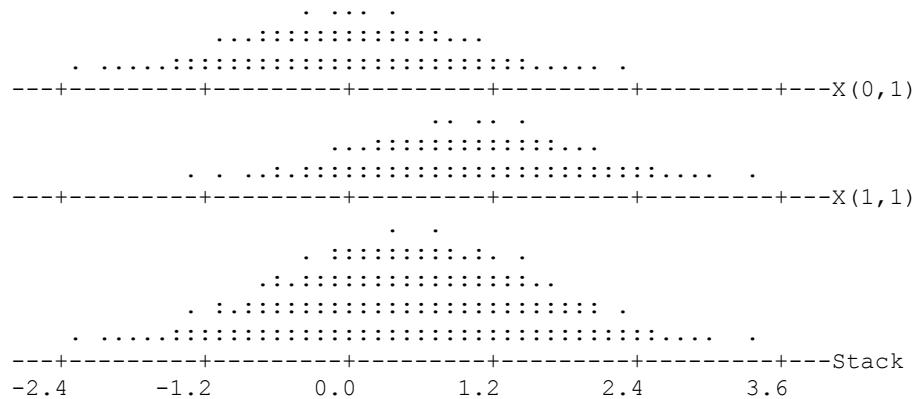
```
                              . ... .
                         ...::::::::::::::...
                 . .....::::::::::::::::::::::::..... .
            ---+---------+---------+---------+---------+---------+---X(0,1)
                                   .. .. .
                              ...::::::::::::...
                      . . ..:.:::::::::::::::::::::::.... .
            ---+---------+---------+---------+---------+---------+---X(1,1)
                                  .  .
                          . :::::::::.:. .
                          .:.::::::::::::::::..
                     . :.:::::::::::::::::::::::::: .
                . .....::::::::::::::::::::::::::::::::.... .
            ---+---------+---------+---------+---------+---------+---Stack
             -2.4      -1.2       0.0       1.2       2.4       3.6

MTB > brief 1
MTB > regress c4 1 predictor c5
```
Regression Analysis

```
The regression equation is
Stack = - 1.00 + 1.00 Subscript

Predictor       Coef        StDev          T          P
Constant     -1.0000       0.2169      -4.61      0.000
Subscrip      1.0000       0.1372       7.29      0.000

S = 0.9652     R-Sq = 21.3%     R-Sq(adj) = 20.9%

Analysis of Variance

Source        DF          SS          MS          F          P
Regression     1      49.500      49.500      53.14      0.000
Error        196     182.588       0.932
Total        197     232.088
```

```
MTB > Stack c7 c3 c8;
SUBC > subscripts c9.


MTB > desc c7 c3 c8
Descriptive Statistics

Variable          N      Mean    Median   Tr Mean    StDev   SE Mean
X(-1,1)          99   -1.0000   -1.0000   -1.0000   0.9652    0.0970
X(1,1)           99    1.0000    1.0000    1.0000   0.9652    0.0970
Stack2          198    0.0000    0.0000    0.0000   1.3899    0.0988

Variable        Min       Max        Q1        Q3
X(-1,1)      -3.3263    1.3263   -1.6745   -0.3255
X(1,1)       -1.3263    3.3263    0.3255    1.6745
Stack2       -3.3263    3.3263   -1.0511    1.0511


MTB > dotplot c7 c3 c8;
SUBC> same.
Character Dotplot


                            . .. ..
                        ...::::::::::::::...
          .   ....:::::::::::::::::::::::::::.:.. . .
         -------+---------+---------+---------+---------+--------X(-1,1)
                                        .. .. .
                                 ...::::::::::::::...
                      . . ..:.::::::::::::::::::::::::.... .
         -------+---------+---------+---------+---------+--------X(1,1)
                       ...: :::.::::::::.::: :...
                      ...:::::::::::::::::::::::::::...
          .  ....:::::::::::::::::::::::::::::::::::::::.... .
         -------+---------+---------+---------+---------+--------Stack2
             -2.4      -1.2       0.0       1.2       2.4       3.6


MTB > Regress 'Stack2' 1 'Subscr2';
SUBC>   Constant.

Regression Analysis


The regression equation is
Stack2 = - 3.00 + 2.00 Subscr2

Predictor      Coef      StDev         T        P
Constant    -3.0000     0.2169    -13.83    0.000
Subscr2      2.0000     0.1372     14.58    0.000

S = 0.9652     R-Sq = 52.0%     R-Sq(adj) = 51.8%

Analysis of Variance

Source       DF        SS        MS        F        P
Regression    1    198.00    198.00   212.54    0.000
Error       196    182.59      0.93
Total       197    380.59
```
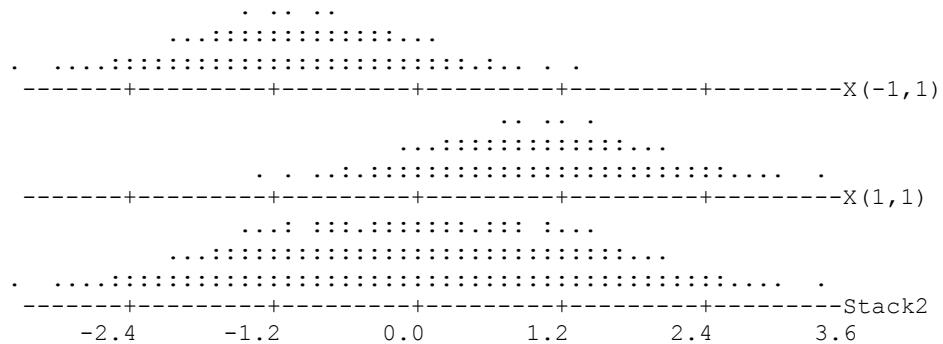
COMPARE MEDIAN OVERLAP AND MISCLASSIFICATION %.

```
MTB > Set c1
DATA>   1( .01 : .5 / .01 )1
DATA>   End.
MTB > InvCDF c1 c2;
SUBC>   Normal 0.0 1.0.
MTB > let c3 = .707*c2
MTB > CDF c3 c4;
SUBC>   Normal 0.0 1.0.


MTB > regress c4 1 c1;
SUBC> predict .5.
```

Regression Analysis

The regression equation is
%Miss = 0.0920 + 0.851 MDN_OVLP

Predictor	Coef	StDev	T	P
Constant	0.092008	0.004009	22.95	0.000
MDN_OVLP	0.85067	0.01368	62.17	0.000

S = 0.01396   R-Sq = 98.8%   R-Sq(adj) = 98.7%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	0.75348	0.75348	3865.15	0.000
Error	48	0.00936	0.00019		
Total	49	0.76284			

Unusual Observations

Obs	MDN_OVLP	%Miss	Fit	StDev Fit	Residual	St Resid
1	0.010	0.05001	0.10051	0.00389	-0.05050	-3.77R
2	0.020	0.07325	0.10902	0.00377	-0.03577	-2.66R

R denotes an observation with a large standardized residual

Fit	StDev Fit	95.0% CI	95.0% PI
0.51734	0.00389	(0.50952, 0.52516)	(0.48819, 0.54649)

```
MTB > desc c1 c12

Variable        N      Mean    Median   Tr Mean    StDev   SE Mean
MDN_OVLP       50    0.2550    0.2550    0.2550   0.1458    0.0206
R-sq           50    0.1504    0.0979    0.1355   0.1536    0.0217


Variable      Min       Max        Q1        Q3
MDN_OVLP   0.0100    0.5000    0.1275    0.3825
R-sq       0.0000    0.5750    0.0219    0.2448

MTB > regress c12 1 c4
The regression equation is
R-sq = 0.518 - 1.19 Miss


Predictor       Coef        StDev          T         P
Constant     0.51788      0.01525      33.97     0.000
Miss        -1.18942      0.04582     -25.96     0.000


S = 0.04002     R-Sq = 93.3%     R-Sq(adj) = 93.2%


Analysis of Variance
Source       DF          SS          MS          F          P
Regression    1      1.0792      1.0792     673.78      0.000
Error        48      0.0769      0.0016
Total        49      1.1561


Unusual Observations
Obs     Miss        R-sq        Fit   StDev Fit    Residual    St Resid
  1    0.050     0.57501    0.45840     0.01314     0.11661       3.08R
  2    0.073     0.51326    0.43076     0.01219     0.08250       2.16R
R denotes an observation with a large standardized residual

MTB > regress c4 1 c12
The regression equation is
Miss = 0.427 - 0.785 R-sq


Predictor       Coef        StDev          T         P
Constant    0.426995     0.006467      66.02     0.000
R-sq       -0.78483      0.03024      -25.96     0.000


S = 0.03251     R-Sq = 93.3%     R-Sq(adj) = 93.2%


Analysis of Variance
Source       DF          SS          MS          F          P
Regression    1     0.71211     0.71211     673.78      0.000
Error        48     0.05073     0.00106
Total        49     0.76284


Unusual Observations
Obs      R-sq        Miss        Fit   StDev Fit    Residual    St Resid
  1     0.575     0.05001   -0.02429     0.01364     0.07430      2.52RX
  2     0.513     0.07325    0.02417     0.01189     0.04908      1.62 X
 49     0.000     0.49293    0.42687     0.00646     0.06606      2.07R
 50     0.000     0.50000    0.42699     0.00647     0.07301      2.29R
R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

```
MTB > let c5 = c4 - c1
MTB > let c6 = c4/c1
MTB > let c7 = (c4-c1)/c1
MTB > print c1 c2 c4 c5 c6 c7
 Row  MDN_OVLP         Z        Miss         Diff      Ratio    RelDiff
   1      0.01  -2.32635   0.050013    0.0400129    5.00129    4.00129
   2      0.02  -2.05375   0.073251    0.0532508    3.66254    2.66254
   3      0.03  -1.88079   0.091805    0.0618052    3.06017    2.06017
   4      0.04  -1.75069   0.107907    0.0679072    2.69768    1.69768
   5      0.05  -1.64485   0.122433    0.0724328    2.44866    1.44866
   6      0.06  -1.55477   0.135835    0.0758350    2.26392    1.26392
   7      0.07  -1.47579   0.148385    0.0783852    2.11979    1.11979
   8      0.08  -1.40507   0.160261    0.0802611    2.00326    1.00326
   9      0.09  -1.34076   0.171587    0.0815867    1.90652    0.90652
  10      0.10  -1.28155   0.182453    0.0824528    1.82453    0.82453
  11      0.11  -1.22653   0.192928    0.0829284    1.75389    0.75389
  12      0.12  -1.17499   0.203067    0.0830671    1.69223    0.69223
  13      0.13  -1.12639   0.212912    0.0829118    1.63778    0.63778
  14      0.14  -1.08032   0.222497    0.0824974    1.58927    0.58927
  15      0.15  -1.03643   0.231853    0.0818528    1.54569    0.54569
  16      0.16  -0.99446   0.241002    0.0810024    1.50626    0.50626
  17      0.17  -0.95417   0.249967    0.0799666    1.47039    0.47039
  18      0.18  -0.91537   0.258763    0.0787631    1.43757    0.43757
  19      0.19  -0.87790   0.267407    0.0774075    1.40741    0.40741
  20      0.20  -0.84162   0.275913    0.0759129    1.37956    0.37956
  21      0.21  -0.80642   0.284291    0.0742914    1.35377    0.35377
  22      0.22  -0.77219   0.292553    0.0725534    1.32979    0.32979
  23      0.23  -0.73885   0.300708    0.0707082    1.30743    0.30743
  24      0.24  -0.70630   0.308764    0.0687643    1.28652    0.28652
  25      0.25  -0.67449   0.316729    0.0667294    1.26692    0.26692
  26      0.26  -0.64335   0.324610    0.0646103    1.24850    0.24850
  27      0.27  -0.61281   0.332413    0.0624134    1.23116    0.23116
  28      0.28  -0.58284   0.340145    0.0601445    1.21480    0.21480
  29      0.29  -0.55338   0.347809    0.0578089    1.19934    0.19934
  30      0.30  -0.52440   0.355412    0.0554115    1.18471    0.18471
  31      0.31  -0.49585   0.362957    0.0529570    1.17083    0.17083
  32      0.32  -0.46770   0.370450    0.0504496    1.15765    0.15765
  33      0.33  -0.43991   0.377893    0.0478933    1.14513    0.14513
  34      0.34  -0.41246   0.385292    0.0452920    1.13321    0.13321
  35      0.35  -0.38532   0.392649    0.0426490    1.12185    0.12185
  36      0.36  -0.35846   0.399968    0.0399679    1.11102    0.11102
  37      0.37  -0.33185   0.407252    0.0372518    1.10068    0.10068
  38      0.38  -0.30548   0.414504    0.0345037    1.09080    0.09080
  39      0.39  -0.27932   0.421727    0.0317266    1.08135    0.08135
  40      0.40  -0.25335   0.428923    0.0289232    1.07231    0.07231
  41      0.41  -0.22754   0.436096    0.0260962    1.06365    0.06365
  42      0.42  -0.20189   0.443248    0.0232483    1.05535    0.05535
  43      0.43  -0.17637   0.450382    0.0203819    1.04740    0.04740
  44      0.44  -0.15097   0.457499    0.0174995    1.03977    0.03977
  45      0.45  -0.12566   0.464603    0.0146035    1.03245    0.03245
  46      0.46  -0.10043   0.471696    0.0116962    1.02543    0.02543
  47      0.47  -0.07527   0.478780    0.0087800    1.01868    0.01868
  48      0.48  -0.05015   0.485857    0.0058570    1.01220    0.01220
  49      0.49  -0.02507   0.492930    0.0029296    1.00598    0.00598
  50      0.50   0.00000   0.500000    0.0000000    1.00000    0.00000
```

```
MTB > let c12 = 1/ (1 + 1/(C2**2/4))    # R-sq using Z.
MTB > let c13 = .401 - .982*c1          # Regress R2 on Mdn_Ovlp
MTB > let c14 = (.5-c4)/.5              # % reduction in misclassify
MTB > Print c1 c2 c4 c12 c13 c14
```

Row	MDN_OVLP	Z	Miss	R-sq	Regr(R2)	Reduction
1	0.01	-2.32635	0.050013	0.575006	0.39118	0.899974
2	0.02	-2.05375	0.073251	0.513257	0.38136	0.853498
3	0.03	-1.88079	0.091805	0.469312	0.37154	0.816390
4	0.04	-1.75069	0.107907	0.433821	0.36172	0.784186
5	0.05	-1.64485	0.122433	0.403479	0.35190	0.755134
6	0.06	-1.55477	0.135835	0.376687	0.34208	0.728330
7	0.07	-1.47579	0.148385	0.352537	0.33226	0.703230
8	0.08	-1.40507	0.160261	0.330457	0.32244	0.679478
9	0.09	-1.34076	0.171587	0.310062	0.31262	0.656827
10	0.10	-1.28155	0.182453	0.291079	0.30280	0.635094
11	0.11	-1.22653	0.192928	0.273305	0.29298	0.614143
12	0.12	-1.17499	0.203067	0.256588	0.28316	0.593866
13	0.13	-1.12639	0.212912	0.240808	0.27334	0.574176
14	0.14	-1.08032	0.222497	0.225870	0.26352	0.555005
15	0.15	-1.03643	0.231853	0.211697	0.25370	0.536294
16	0.16	-0.99446	0.241002	0.198228	0.24388	0.517995
17	0.17	-0.95417	0.249967	0.185408	0.23406	0.500067
18	0.18	-0.91537	0.258763	0.173194	0.22424	0.482474
19	0.19	-0.87790	0.267407	0.161549	0.21442	0.465185
20	0.20	-0.84162	0.275913	0.150441	0.20460	0.448174
21	0.21	-0.80642	0.284291	0.139843	0.19478	0.431417
22	0.22	-0.77219	0.292553	0.129731	0.18496	0.414893
23	0.23	-0.73885	0.300708	0.120085	0.17514	0.398584
24	0.24	-0.70630	0.308764	0.110887	0.16532	0.382471
25	0.25	-0.67449	0.316729	0.102120	0.15550	0.366541
26	0.26	-0.64335	0.324610	0.093771	0.14568	0.350779
27	0.27	-0.61281	0.332413	0.085827	0.13586	0.335173
28	0.28	-0.58284	0.340145	0.078278	0.12604	0.319711
29	0.29	-0.55338	0.347809	0.071114	0.11622	0.304382
30	0.30	-0.52440	0.355412	0.064327	0.10640	0.289177
31	0.31	-0.49585	0.362957	0.057908	0.09658	0.274086
32	0.32	-0.46770	0.370450	0.051850	0.08676	0.259101
33	0.33	-0.43991	0.377893	0.046148	0.07694	0.244213
34	0.34	-0.41246	0.385292	0.040796	0.06712	0.229416
35	0.35	-0.38532	0.392649	0.035790	0.05730	0.214702
36	0.36	-0.35846	0.399968	0.031123	0.04748	0.200064
37	0.37	-0.33185	0.407252	0.026794	0.03766	0.185496
38	0.38	-0.30548	0.414504	0.022798	0.02784	0.170993
39	0.39	-0.27932	0.421727	0.019132	0.01802	0.156547
40	0.40	-0.25335	0.428923	0.015793	0.00820	0.142154
41	0.41	-0.22754	0.436096	0.012779	-0.00162	0.127808
42	0.42	-0.20189	0.443248	0.010087	-0.01144	0.113503
43	0.43	-0.17637	0.450382	0.007717	-0.02126	0.099236
44	0.44	-0.15097	0.457499	0.005666	-0.03108	0.085001
45	0.45	-0.12566	0.464603	0.003932	-0.04090	0.070793
46	0.46	-0.10043	0.471696	0.002515	-0.05072	0.056608
47	0.47	-0.07527	0.478780	0.001414	-0.06054	0.042440
48	0.48	-0.05015	0.485857	0.000628	-0.07036	0.028286
49	0.49	-0.02507	0.492930	0.000157	-0.08018	0.014141
50	0.50	0.00000	0.500000	0.000000	-0.09000	0.000000

```
MTB > corr c1-c4 c12-c14

           MDN_OVLP        Z         Zp       Miss      R-sq    Regr(R2)
Z           0.966
Zp          0.966     1.000
Miss        0.994     0.988     0.988
R-sq       -0.932    -0.991    -0.991    -0.966
Regr(R2)   -1.000    -0.966    -0.966    -0.994     0.932
Reductio   -0.994    -0.988    -0.988    -1.000     0.966     0.994
```