

Correlation

Review and Extension

First of all, what is it?

- Pearson r
- Standardized covariance
 - Covariance, like variance is a statistic on its own
 - Measure the degree to which two variables *covary*
 - This is a standardized version putting it on a -1 to +1 scale for easier interpretation
- Pearson's r is used as a descriptive statistic in initial data examination, on its own it is no more informative than means or standard deviations and other descriptive measures
- However, many multivariate methods use a correlation matrix as the data input, rather than raw variable values
- In either case the measure of association is an important basis for inferential measures¹ and as such any problems with it will cause the procedures based on it to fail

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

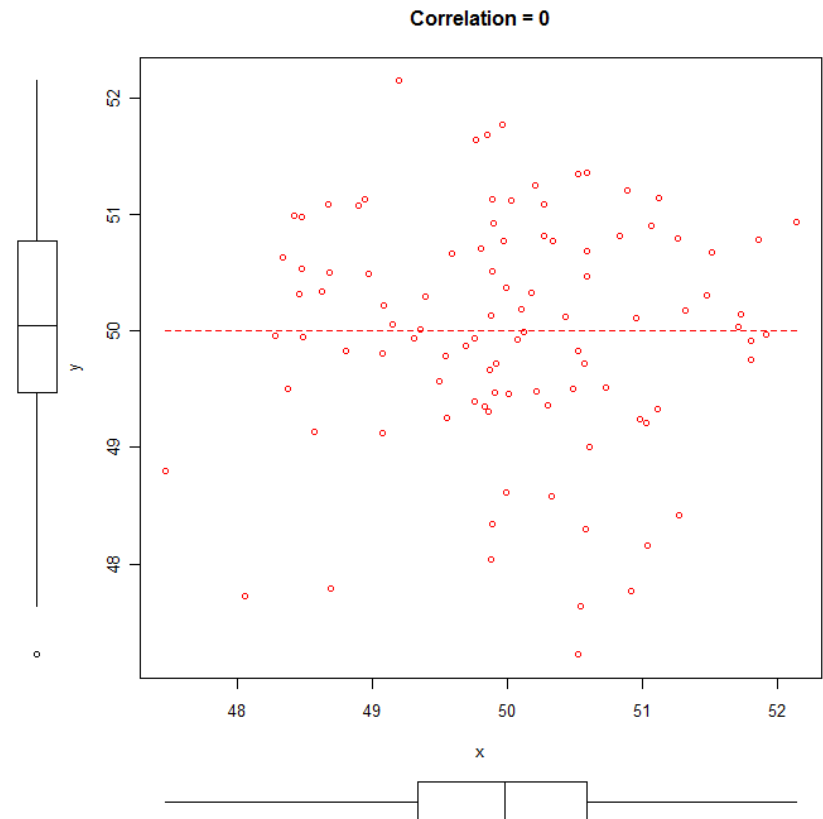
$$r_{xy} = \frac{\text{cov}(x, y)}{s_x s_y}$$

Factors Affecting the Pearson r

- Magnitude of residuals
- Slope of the regression line
- Outliers
- Curvature
- Restriction of range
 - In terms of reliability and measurement, restricting the range will attenuate r
 - In terms of outliers, restriction of range may increase or decrease depending on the nature of the outlier
- Context
 - Subsamples
 - Individuals

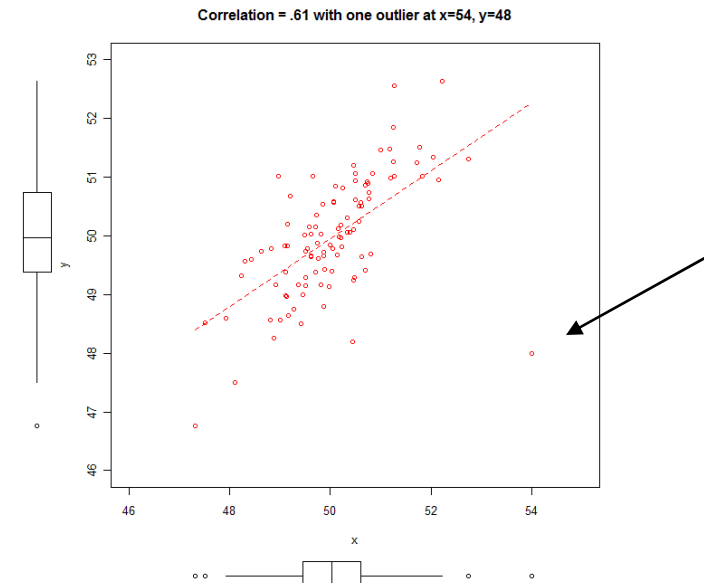
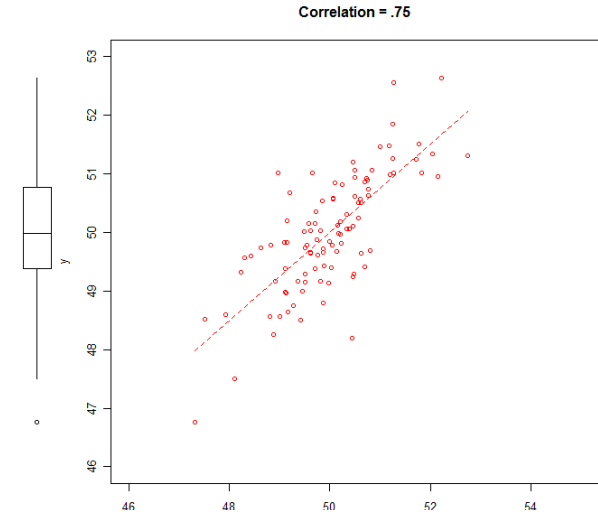
Factors Affecting the Pearson r

- Magnitude of residuals
 - Errors in prediction
 $Y_{\text{pred}} - Y_{\text{obs}}$
 - More scatter about the regression line, less association between variables
- Slope of the regression line
 - All else being equal, as slope approaches zero



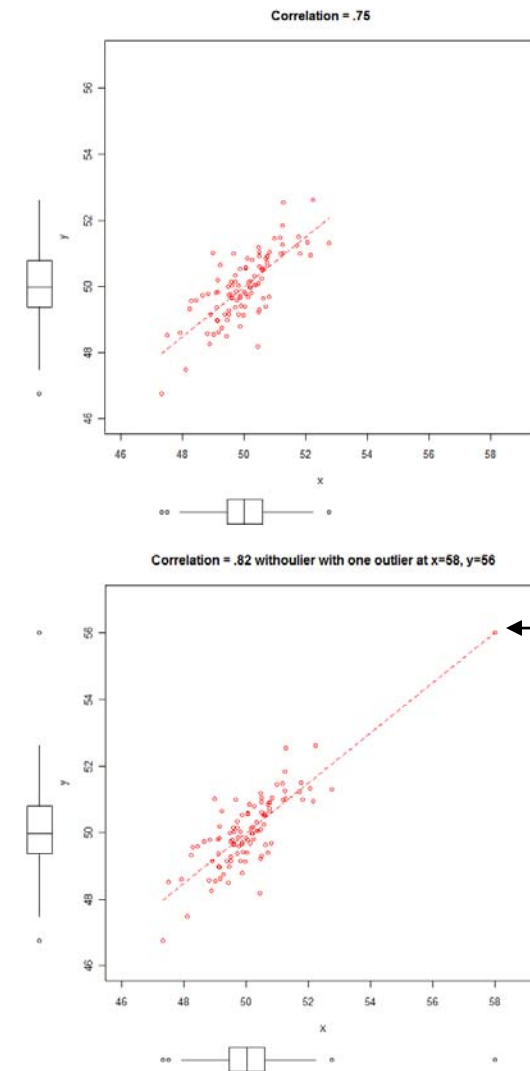
Factors Affecting the Pearson r

- Outliers
- The example to the right shows the huge drop in the correlation due to an outlier
- Despite $N = 100$, just adding one point (which was still within the range of original DV values) completely distorted the true correlation



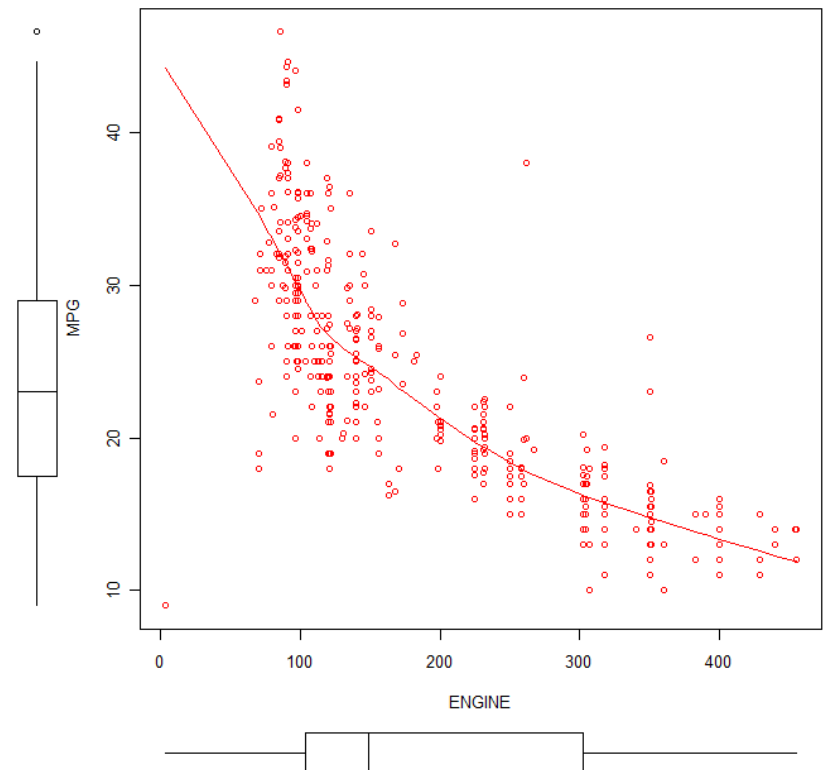
Factors Affecting the Pearson r

- Outliers can work in the opposite direction, distorting the nature of the relationship in optimistic fashion
- Again, same data as starting point (N=100), one point was added directly on the least squares line
- However it is clearly an outlier on both variables



Factors Affecting the Pearson r

- Curvature
- Curvilinear relationships will have an adverse effect on a measure designed to find a strictly linear relationship
- While one might consider data transformation, such a course may actually hamper interpretability
- Smoothing methods are available to help¹ and should be a basic part of initial data examination of variables and in the inferential model testing stage as well
 - E.g. Loess (locally weighted scatterplot smoothing)



Factors Affecting the Pearson r

- Restriction of range
- Intuition: If you wanted to measure the relationship of age to self-esteem among adolescents and young adults (13-25), assuming the relationship was a notable one, which do you think is likely to reveal this?
 - Correlation of age in months vs. an SE scale ranging from 0-30
 - Or
 - Correlation between age young-old vs. SE low-high using arbitrary median splits¹
- However in the case of our first outlier, restricting the range of X values would return the correlation back to normal

Factors Affecting the Pearson r

- Context
- Another reason correlations in isolation are not very informative is that contextual factors are not taken into account
- Relationships can change depending on gender, SES, geographical regions etc.
- Thinking in this manner gets us into notions of model comparison, interactions, multilevel modeling, Bayesian methods etc., as intro text book approaches won't

Other uses of r

- Squaring the r tells us the *shared variance* of two variables, or if we consider one the model dependent variable, the amount of *variance accounted for in the DV by the predictor*
- Additionally, if we standardize the data r is the slope in the simple regression situation, and allows us to get standardized regression coefficients in multiple regression¹

Old school and modern approaches to understanding relationships

- Changing data values
 - Ranking
 - Winsorizing
- Downweighting extreme endpoints
 - M-estimator methods
- Bivariate approaches

Changing data values

- *Ranks*
 - A common robust approach to regression involves transforming the entire dataset into ranks
 - $81,93,67,52, 56 \rightarrow 4,5,3,1,2$
 - This maintains the ordinality of the data while reeling in outliers
 - The correlation is thus between two ranked variables
 - Examples: Spearman's rho, Kendall's tau
- *Winsorizing*
 - Also takes a transformation approach but leaves the bulk of the data unchanged
 - Transform X% of the tails of the data to the last value before that cutoff
 - $81,93,67,52,56 \rightarrow 52,56,67,81,93 \rightarrow 56,56,67,81,81$
 - As you are familiar with trimming, it is the same process, except instead of trimming the values we change a certain percentage of the most extreme values (not necessarily outliers) to less extreme ones
 - This process is done for both variables of interest, and then the correlation is calculated in the same manner as always

Other correlations for certain data situations¹

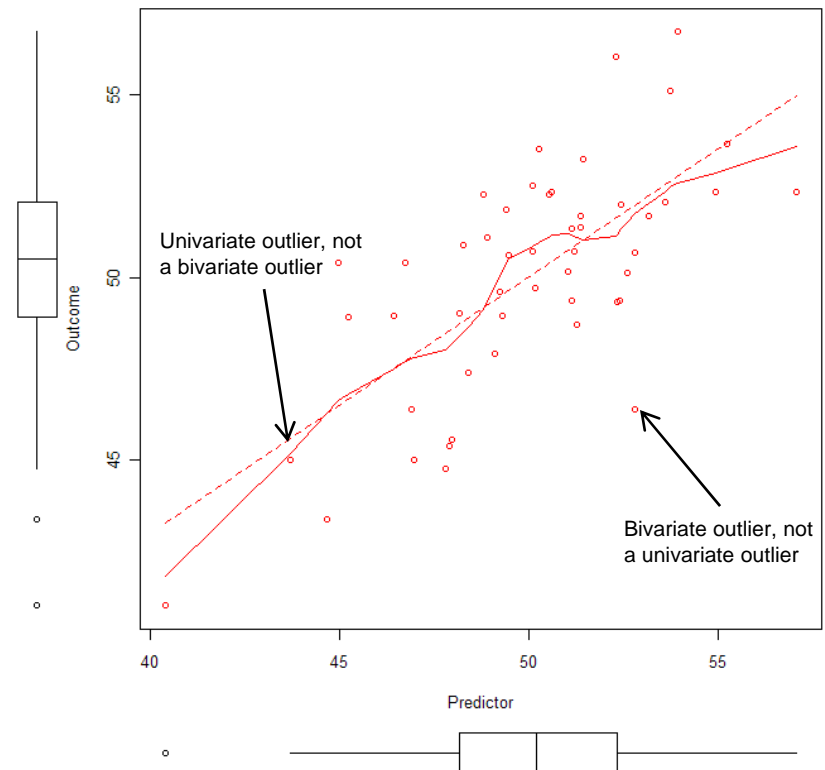
- Underlying continuums
- Biserial correlation
 - Used with a continuous and categorical in which the latter is assumed to represent an underlying continuum (but you aren't going to dichotomous continuous variables are you?)
 - Compared to the point-biserial which you are familiar with from a independent sample t-test effect size, biserial $>$ r_{pb}
- Tetrachoric and polychoric
 - Polychoric is the more general of the two: a correlation is calculated on ordinal data (e.g. Likert) based on an assumed continuum
 - Tetrachoric is used when both variables are dichotomous
 - Both are useful for SEM type applications, and most useful SEM software will do this

Modern approaches

- Methods related to M-estimators
- These are similar to the trimming/Winsorizing approach except that extreme cases are determined by the data, as opposed to picking a percentage without regard to the actual data, and downweighted according to how extreme they are

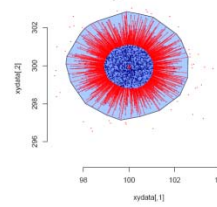
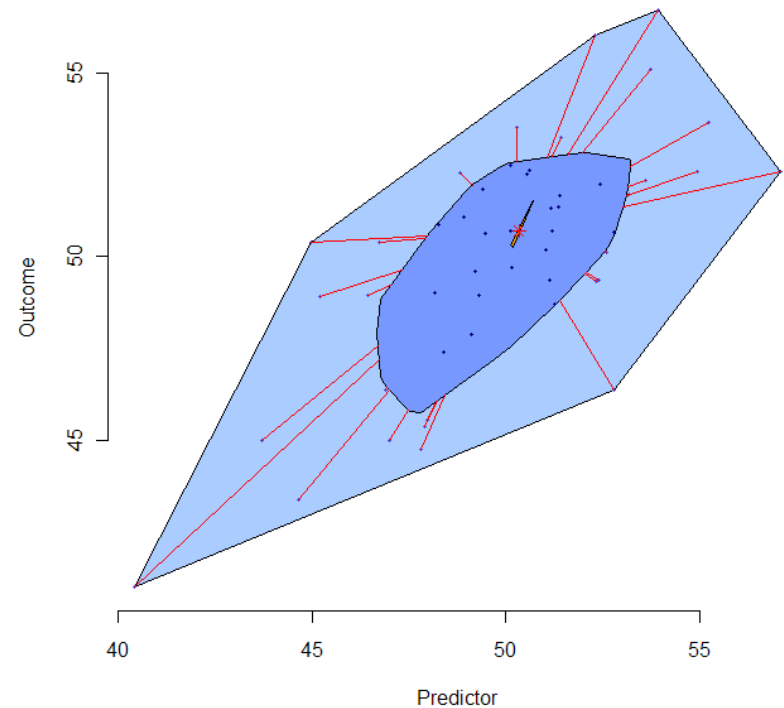
But what about the model?

- The problem with these methods is that they are not taking into account the bivariate nature of the data.
- An extreme value for one or even both variables may not be at all with regard to the relationship between the two, and non-outliers on either variable might



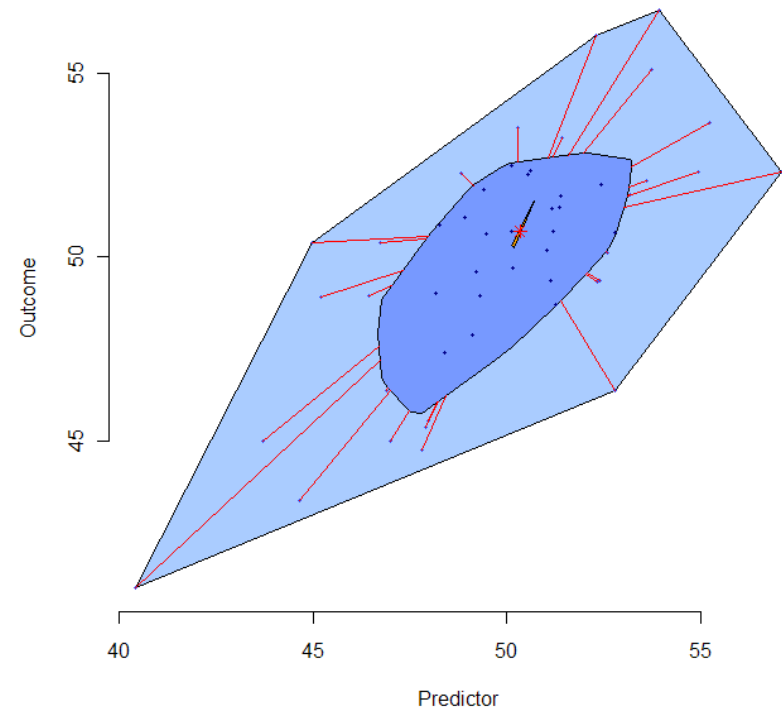
Examples of robust correlation

- At right is a bagplot or bivariate boxplot¹
- There is the middle 50% of values- the “bag”
 - The mark in the bag represents the bivariate median
- Anything outside the light blue “fence” (determined by researcher, in this case the bag increased by a factor of 3) would be considered an outlier
- If all points were on a line, it would be a boxplot
- Below is a bivariate normal distribution of independent variables



Examples of robust correlation

- Minimum volume ellipsoid estimator
 - MVE
- Takes a similar approach and finds the 50% of data that would create the smallest bag, and computes the correlation from those data points



Examples of robust correlation

- Minimum Covariance Determinant Estimator
 - MCD
- The underlying basis of the MCD requires multivariate course knowledge
- However for our purposes it is enough to know that it is similar in nature to the MVE in selecting a ‘best’ 50% of the data in some optimal fashion

- Example R code¹:

```
library(robust)
covRob(Predictor, Outcome, corr=T)
```

```
#Alternate
covRob(mydata, corr=T)
```

- Result

```
Robust Estimate of Correlation:
```

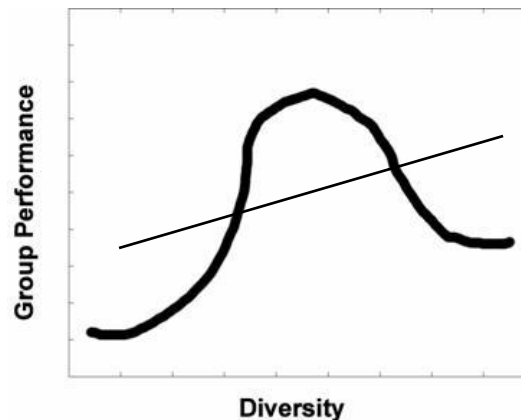
	Predictor	Outcome
Predictor	1.0000000	0.5293861
Outcome	0.5293861	1.0000000

```
Robust Estimate of Location:
```

Predictor	Outcome
50.51780	50.72738

Remaining issues: Curvature

- The fact is that straight lines may not capture the true story
- We may often fail to find noticeable relationships because our r , whichever method of “Pearsonesque” one we choose, is trying to specify a linear relationship
- There may still be a relationship, and a strong one, just more complex



Summary

- Correlation, in terms of Pearson r , gives us a sense of the strength of a linear association between two variables
- One data point can render it a useless measure, as it is not robust to outliers
- Measures which are robust are available, and some take into account the bivariate nature of the data
- However, curvilinear relationships may exist, and we should examine the data to see if alternative explanations are viable