Psychological Variables and Academic Grades

Contents

Overview

The data file 'psy_grades.txt' contains data measuring three psychological variables and four academic variables (grades) for 300 students. The sex of each student and the school they attend was also recorded.

Psychological variables:

- control: locus of control is the degree to which people believe that they have control over the outcome of events in their lives.
- self: self-concept is a collection of beliefs about oneself that includes elements such as academic performance.
- motive: motivation is a measure achievement motivation.

Academic variables:

- english: grade of performance over 1 year.

- history: grade of performance over 1 year.
- maths: grade of performance over 1 year.
- biology: grade of performance over 1 year.

Other variables:

- sex: 0=Male and 1=Female
- school: school 1, 2 or 3
- the row labels are the student ID number

Assumptions: Assume all variables meet MVN and other test assumptions for the purpose of these assessment questions.

Would canonical correlation would be an appropriate form of analysis ?

Canonical correlations are a measure of the strength of the overall linear relationships between the canonical variates for the independent and dependent variables. They represent the bivariate correlation between these two canonical variates (U and V) of each canonical function. Canonical correlations are the square-roots of the eigenvalues.

In Canonical Correlation Analysis (CCA) very high or very low correlations can cause misleading results or failure of the analysis.

I will standardise the variables and produce 3 separate pairwise correlation matrices: i. correlation between the 3 psychological variables; ii. correlation between the 4 academic variables; iii. correlation between the 3 psychological variables and the 4 academic variables .

```
#read and preview data
pg <- read.table("C:/Users/arjom/OneDrive/Desktop/R Working Directory/psy_grades.txt" , header = TRUE)
str(pg)
```

```
## 'data.frame':    300 obs. of  9 variables:
##  $ control: num  -0.84 -0.38 0.89 0.71 -0.64 1.11 0.06 -0.91 0.45 0 ...
##  $ self   : num  -0.24 -0.47 0.59 0.28 0.03 0.9 0.03 -0.59 0.03 0.03 ...
##  $ motive : num  1 0.67 0.67 0.67 1 0.33 0.67 0.67 1 0.67 ...
##  $ english: num  54.8 62.7 60.6 62.7 41.6 62.7 41.6 44.2 62.7 62.7 ...
##  $ history: num  64.5 43.7 56.7 56.7 46.3 64.5 39.1 39.1 51.5 64.5 ...
##  $ maths  : num  44.5 44.7 70.5 54.7 38.4 61.4 56.3 46.3 54.4 38.3 ...
##  $ biology: num  52.6 52.6 58 58 36.3 58 45 36.3 49.8 55.8 ...
##  $ sex    : int  1 1 0 0 1 1 0 0 1 1 ...
##  $ school : int  2 2 1 2 3 1 2 3 2 2 ...
```

```
#standardised mean=0 and stdev=1
pgst<- scale(pg[1:7])[,]
#check correlations
psych <- pgst[, 1:3]
acad <- pgst[, 4:7]
cor(psych,psych)
```

```
##           control      self    motive
## control 1.0000000 0.1757018 0.2477818
## self    0.1757018 1.0000000 0.2601119
## motive  0.2477818 0.2601119 1.0000000
```

```
cor(psych,acad)
```

```
##            english    history     maths   biology
## control 0.41014965 0.36333123 0.3841518 0.3413628
## self    0.06604255 0.06904613 0.1241400 0.1153015
## motive  0.21720996 0.29034813 0.2139463 0.1665264
```

```
cor(acad,acad)
```

```
##           english   history     maths   biology
## english 1.0000000 0.5939160 0.6936508 0.6923142
## history 0.5939160 1.0000000 0.6137526 0.5478822
## maths   0.6936508 0.6137526 1.0000000 0.6477564
## biology 0.6923142 0.5478822 0.6477564 1.0000000
```

The psychological variables have both only positive correlations which are all quite low, ranging from (0.18) to (0.26). Motive is most strongly related with the other two. The academic variables are also all positively correlated however the correlations are all moderately strong between 0.55 and 0.69. The strongest correlations are of English with maths and biology, both with a correlation of 0.69. The psychological and academic correlations are very low to weakly moderate (0.07 to 0.41), with the strongest correlation occurring across the groups between English and control. In CCA very high or very low correlations can cause misleading results or failure of the analysis. In this case I think there are enough moderate correlations for the analysis to be worth a try, however the low correlations among the psychological variables may cause some issues with some methods.

Performing Canonical Correlation Analysis

I will perform a CCA on this data set for the standardised variables X1 to X3 (control, self, motive) and Y1 to Y4 (english, history, maths, biology).

```
library(yacca)
pg.cc <- cca(psych, acad)
summary(pg.cc)
```

```
##
## Canonical Correlation Analysis - Summary
##
##
## Canonical Correlations:
##
##       CV 1       CV 2       CV 3
## 0.47846007 0.15243019 0.08380162
##
## Shared Variance on Each Canonical Variate:
##
##        CV 1        CV 2        CV 3
## 0.228924041 0.023234963 0.007022712
##
## Bartlett's Chi-Squared Test:
##
```

```
##           rho^2      Chisq df    Pr(>X)
## CV 1  0.2289240 85.7048834 12 3.331e-13 ***
## CV 2  0.0232350  9.0142082  6    0.1728
## CV 3  0.0070227  2.0790087  2    0.3536
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Canonical Variate Coefficients:
##
##  X Vars:
##                CV 1       CV 2       CV 3
## control -0.83987825 -0.3838375  0.4774617
## self     0.05860289 -0.7237370 -0.7488162
## motive  -0.39506726  0.8415253 -0.5090129
##
##  Y Vars:
##                 CV 1       CV 2       CV 3
## english -0.451918372  0.2457541  1.4933987
## history -0.447667192  1.0749128 -0.5893716
## maths   -0.242968497 -0.6860952 -0.6391102
## biology -0.007077361 -0.8023382 -0.3938551
##
##
## Structural Correlations (Loadings):
##
##  X Vars:
##               CV 1       CV 2       CV 3
## control -0.9274721 -0.3024847  0.2197693
## self    -0.1917269 -0.5722871 -0.7973257
## motive  -0.5879305  0.5581647 -0.5854826
##
##  Y Vars:
##                CV 1       CV 2        CV 3
## english -0.8912301 -0.1472186  0.42737069
## history -0.8690688  0.3601905 -0.31045997
## maths   -0.8357833 -0.3756169 -0.22006348
## biology -0.7226002 -0.4876961 -0.09684795
##
##
## Fractional Variance Deposition on Canonical Variates:
##
##  X Vars:
##              CV 1       CV 2       CV 3
## control 0.8602045 0.09149697 0.04829852
## self    0.0367592 0.32751254 0.63572825
## motive  0.3456623 0.31154785 0.34278986
##
##  Y Vars:
##              CV 1       CV 2        CV 3
## english 0.7942911 0.02167333 0.182645703
## history 0.7552806 0.12973722 0.096385394
## maths   0.6985338 0.14108806 0.048427935
## biology 0.5221510 0.23784746 0.009379526
```

```
##
##
## Canonical Communalities (Fraction of Total Variance
## Explained for Each Variable, Within Sets):
##
##  X Vars:
## control    self  motive
##       1       1       1
##
##  Y Vars:
##   english   history     maths   biology
## 0.9986102 0.9814033 0.8880498 0.7693780
##
##
## Canonical Variate Adequacies (Fraction of Total Variance
## Explained by Each CV, Within Sets):
##
##
##  X Vars:
##      CV 1      CV 2      CV 3
## 0.4142087 0.2435191 0.3422722
##
##  Y Vars:
##       CV 1       CV 2       CV 3
## 0.69256414 0.13258652 0.08420964
##
##
## Redundancy Coefficients (Fraction of Total Variance
## Explained by Each CV, Across Sets):
##
##
##  X | Y:
##        CV 1        CV 2        CV 3
## 0.094822322 0.005658158 0.002403679
##
##  Y | X:
##        CV 1        CV 2        CV 3
## 0.158544582 0.003080643 0.000591380
##
##
## Aggregate Redundancy Coefficients (Total Variance
## Explained by All CVs, Across Sets):
##
##  X | Y: 0.1028842
##  Y | X: 0.1622166
```

```
F.test.cca(pg.cc)
```

```
##
##  F Test for Canonical Correlations (Rao's F Approximation)
##
##           Corr         F    Num df Den df   Pr(>F)
## CV 1  0.478460  7.500598 12.000000  775.5 3.34e-13 ***
## CV 2  0.152430  1.508772  6.000000  588.0   0.1728
```

```
## CV 3  0.083802        NA  2.000000     NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Canonical correlations represent the shared variance between variants which is the variance shared by the linear composites of dependent and independent variables. The first pair of canonical variates is derived to have the highest inter-correlation possible between the two sets of original variables. The second pair of canonical variates is then derived so that it exhibits the maximum relationship between the two sets of variables not accounted for by the first pair of variates. Successive pairs of canonical variates are based on residual variance, and their respective canonical correlations (which reflect the interrelationships between the variates) become smaller as each additional function is extracted. In the canonical correlation analysis, the psychological variables are treated as the X variables and the academic variables as the Y variables. Shared variance is NOT the variance extracted from the sets of variables and therefore does not add up to 1.

Three canonical variates are produced because the minimum number of original variables in either set is 3.

The canonical correlation for the first canonical function (CV1) is 0.48 which indicates that the correlation between the first 2 canonical variants is moderate. The other canonical functions have much weaker correlations.

The first set of canonical variates are moderately correlated (r=0.48) with each other and their shared variance is 0.23. The eigenvalue (shared variance) of 0.23 represents the variance in one canonical variate explained by the other. Shared variance is the variance shared by the linear composites of dependent and independent variables, and NOT the variance extracted from the sets of variables. Therefore, a relatively low canonical correlation may be obtained between two linear composites (canonical variates).

Bartlett's and Rao's compare all X variables with all Y variables. It is a chi-squared test which tests whether at least one of the 'r' canonical correlations is significant. A small chi-squared value leading to a non-significant result would indicate that even the largest canonical correlation can be accounted for by sampling variation.

Both the Chi-square test and Rao's F approximation test indicate that the first canonical function is significant and worth interpreting. The significance suggests that the canonical correlations is greater than what wold be expected due to sampling variation alone and is instead indicative of a true association between the psychological and academic variables. The significance tests rely on non-violation of MVN, which suggests that our assumption of MVN is not an issue here.

Decisions on whether to interpret the canonical functions should be based not only on the significance tests but also considering the shared variance and redundancy coefficients, which are analogous to R2.

The redundancy coefficient is the amount of variance in a canonical variate (dependent or independent) explained by the other canonical variate in the canonical function.

On the first CV, only 9.5% of the variance in the psychological (X group) variables is explained by the academic (Y group) and 15.9% the academic (Y group) is explained by the psychological (X group). Although these values are fairly low, given th e significance test it may still be worthwhile interpreting the loadings on at least the first canonical function.

The sign of all variable loadings are the same (negative) indicating low psychological values are generally associated with low academic values and high values are similarly associated.

The loadings indicate that students with primarily low control perform poorly in all subjects i.e. control gas the highest loading.

Overall Value of the analysis

Decisions on whether to interpret the canonical functions should be based not only on the significance tests but also considering the shared variance and loadings. Where significance tests indicate a significant

relationship the shared variance and loadings may not reveal useful information. That is somewhat true in this case – the redundancy coefficients are low so although and loadings for CV1 are somewhat interesting CV1 itself does not represent a very strong relationship between the two groups of variables.

Relationship between eigen values and canonical correlations.

The shared variances are the squared canonical correlations and are the eigen values. Therefore, the first 2 canonical variates which make up the first canonical function have a shared variance of 0.223.

For each subsequent function the shared variance is smaller (in line with the smaller canonical correlations).

The shared variance is only an indication of the relationship between variates not the variance extracted from the original data.

Why not multiple regression or MANOVA?

CCA allows for multiple dependant and independent variable either continuous or categorical. Multiple regression has one dependent variable and MANOVA has categorical independent variables.

Although, canonical correlation analysis has its own limitations such as : - CCA only identifies linear relationships. - Although CCA itself does not require MVN it is required for significance tests and often hard to test for. - The lack of assumptions needed for CCA makes it very useful test, however, inferences drawn from it should also be treated with more caution. - CCA also only considers correlation and relationships should not be interpreted as causal.

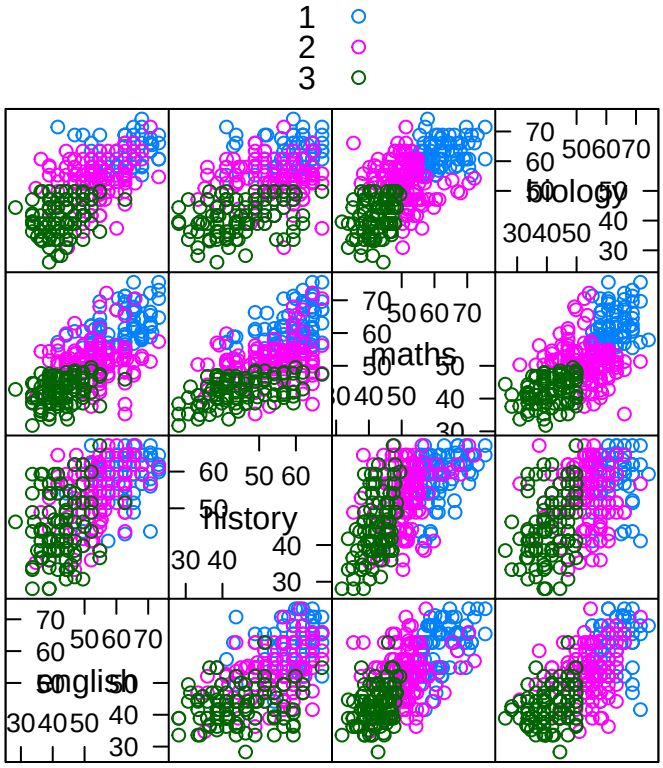
School Prediction

Is it possible to predict the school of a student by their grades across the four Academic variables ?

Pair-wise scatter plots

I use the 'splom' function for all four of the academic variables, distinguishing between schools using colour.

```
library(lattice)
splom(pg[,4:7], groups=pg$school, auto.key=TRUE)
```



Scatter Plot Matrix

- All variable pairs show some positive linear correlation, with the strongest relationships across all students occurring between maths and English and maths and biology.
- The relationship between English and History does not seem to be linearly increasing within school 3 – history grades vary for all levels of English grades and as English increases, History does not.
- The linear relationships within all schools for all variable pairs appear to be much weaker than the correlation patterns across all students.
- School 2 (green) seems consistently distinct from School 1 (blue), however there is overlap with school 2 (pink for both schools 1 and 3.

Training and Test sets

I create the training and test sets using school as a factor, with a 70/30 split and a seed value of 1125.

```
#create training and test sets
pg$school<-as.factor(pg$school)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.6.3
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```r
set.seed(1125)
inTrain <- createDataPartition(y = pg$school, p = .70,list = FALSE)
pgtrain <- pg[ inTrain,]
pgtest <- pg[-inTrain,]
table(pgtrain$school)
```

```
##
##  1  2  3
## 56 88 68
```

```r
table(pgtest$school)
```

```
##
##  1  2  3
## 23 37 28
```

Discriminant Function Analysis (DFA)

I will perfrom the DFA on the training set.

```r
#Linear DA (LDA) = DFA
library(MASS)
(pgtrain.lda<-lda(school~english+history+maths+biology, data=pgtrain))
```

```
## Call:
## lda(school ~ english + history + maths + biology, data = pgtrain)
##
## Prior probabilities of groups:
##         1         2         3
## 0.2641509 0.4150943 0.3207547
##
## Group means:
##    english  history    maths  biology
## 1 62.70536 59.93571 62.91786 62.45179
## 2 53.05341 53.74659 51.35568 53.89091
## 3 41.89118 44.50441 42.59706 40.47794
##
## Coefficients of linear discriminants:
##                  LD1          LD2
## english  0.051184472 -0.001356282
## history -0.004495269  0.056070195
## maths    0.116100984 -0.144097037
## biology  0.092091761  0.092622113
##
## Proportion of trace:
##    LD1    LD2
## 0.9852 0.0148
```

- Why 2 DFs: 4 variables and 3 schools (m) so m-1 is smallest = 3-1=2.
- Prior probabilities = Observed proportion in each school. The prior probabilities for each school are very fairly balanced i.e 26%, 42% and 32% for each of school 1, 2 and 3 respectively.

- The weighting of the original variables on the DF's indicates which variables contribute most to the discrimination between groups on each of the discriminant functions.
- The trace values are the proportions of total between group variance explained by the DF's.
- In this analysis DF1 explains most of the total between group variance (98.5%). This should be interpreted cautiously as if there was not much discrimination between groups in the original data then DF1 is simply explaining 98.5% of not much between group differences.
- None of the original variables have weightings/coefficients greater than |0.14| with most less than |0.1| indicating weak relationships with the discriminant functions.
- DF1 mainly represents maths and to a lesser degree Biology and English. The weighting of history is so low it is effectively not represented.
- This aligns with the earlier interpretation of the scatter plots – strongest relationships between maths, English and biology with the history relationships not appearing linear.

Predict school membership for the test set

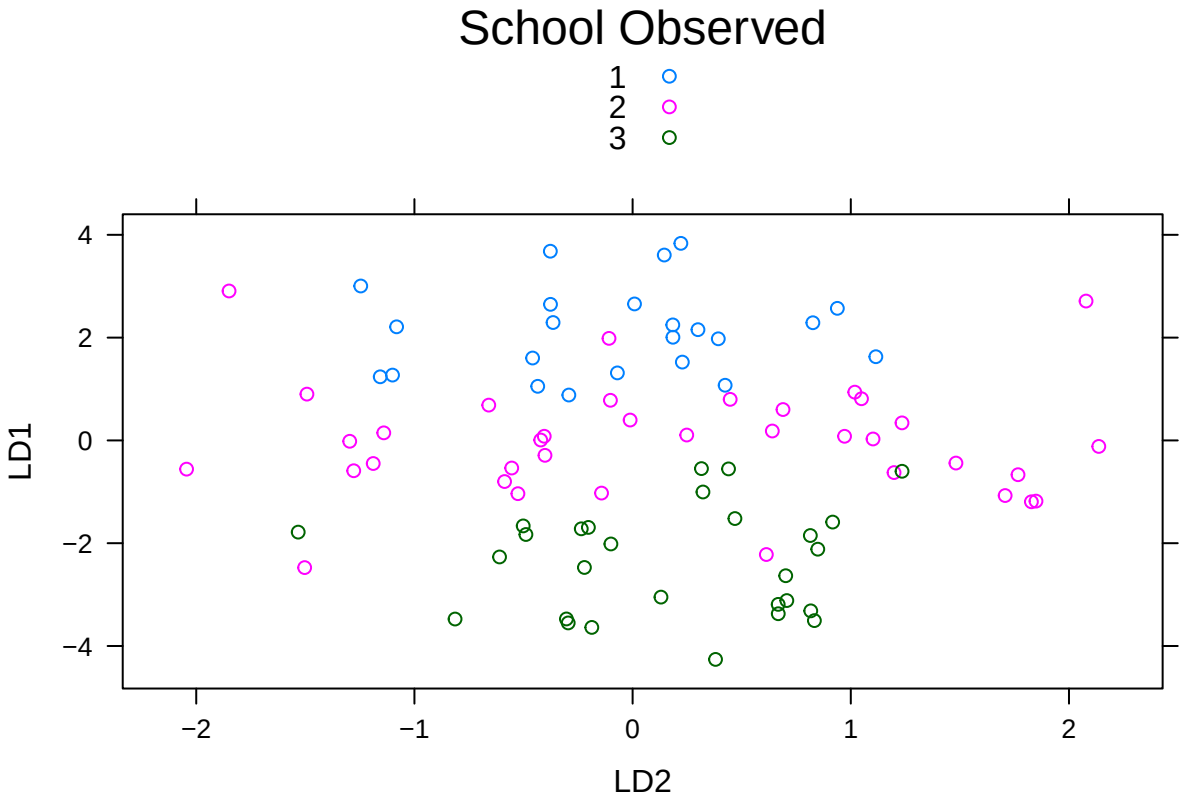
Based on the DFA, I predict school membership for the test set and create and interpret a table showing observed vs predicted for the test set .

Also, I create an x-y plot of the two DFs grouped by the original school labels and another by the predicted school labels.

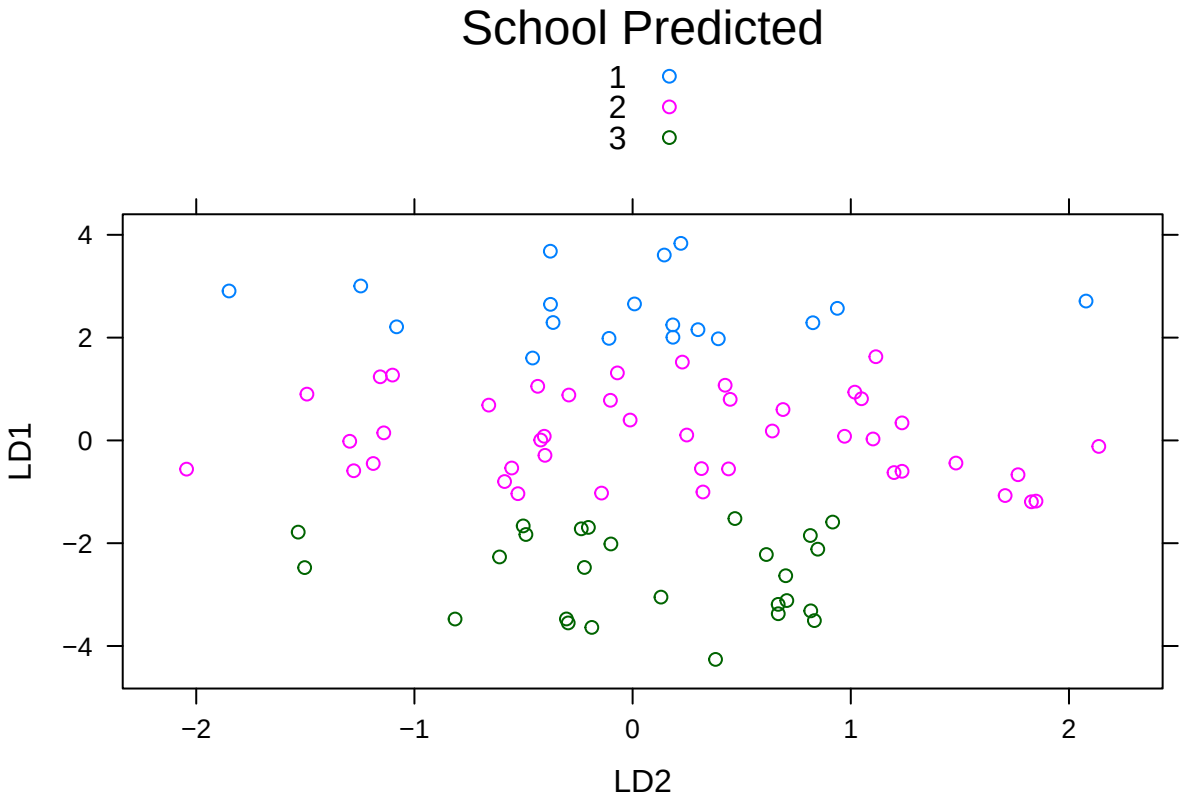
```
school.pred<-predict(pgtrain.lda, pgtest) #predicts school for test set
table(pgtest$school, school.pred$class) #compares true period to predicted for test group
```

```
##
##      1  2  3
##   1 15  8  0
##   2  3 32  2
##   3  0  4 24
```

```
#plot DFs by original school groups
pg.temp<-data.frame(school.pred$x, class=pgtest$school)
xyplot(LD1~LD2, data=pg.temp, groups=class,auto.key=list(title="School Observed", space = "top", cex=1.
```




```
#plot DFs by predicted school groups
pg.temp<-data.frame(school.pred$x, class=school.pred$class)
xyplot(LD1~LD2, data=pg.temp, groups=class,auto.key=list(title="School Predicted", space = "top", cex=1
```



I would expect the misclassification rate for the training set be lower than for the test set as the model is positively biased to fit the data it was built with better than an independent data set.

For rest of the project, let's create a sample :

```
set.seed(24358)
pg_new<- pg[sample(1:nrow(pg), 20, replace=FALSE),]
studentID <- row.names(pg_new)
```

Distance Matrix

I create a distance matrix for students based on the standardised psychological measures using Euclidian distance. I show only the lower triangle of distances and do not include the diagonal zero's. Also, I limit all values to 2 decimal places. In addition, I label the rows and columns with the school number (1, 2 or 3) associated with each student.

```
pg_new_distance<-subset(pg_new[,1:3])
(pg_new_distance.scaled<-data.frame(sapply(pg_new_distance, scale)))
```

```
##       control        self     motive
## 1   0.2281057  0.45967488 -0.7036200
## 2   1.6680648  0.78607125  1.0391726
## 3   0.7529506  0.01087988  0.1807822
## 4   0.2146481 -0.02991967  1.0391726
## 5   0.7529506  1.20766656  1.0391726
```

```
## 6  -1.2925988 -0.02991967 -1.5620104
## 7  -0.9696173  0.39167564 -1.5620104
## 8   0.8471535  1.54766278  1.0391726
## 9   0.1473603 -0.42431528 -1.5620104
## 10 -2.8402185 -0.02991967 -1.5620104
## 11 -0.1621636 -1.55310272  0.1807822
## 12  0.4568842 -0.02991967 -0.7036200
## 13  0.2415632  0.69087231  1.0391726
## 14  1.3316258  0.39167564  1.0391726
## 15  0.2011906 -1.49870332 -0.7036200
## 16 -0.0275880  0.69087231  0.1807822
## 17 -0.1621636  0.81327095  1.0391726
## 18 -0.6600934 -2.65469046  0.1807822
## 19 -0.9696173 -0.70991210  0.1807822
## 20  0.2415632 -0.02991967  0.1807822
```

```
(distpsy <-dist(pg_new_distance.scaled, method="euclidian", diag = FALSE, upper = FALSE))
```

```
##           1         2         3         4         5         6         7
## 2  2.2841504
## 3  1.1220724 1.4748524
## 4  1.8103066 1.6668117 1.0140356
## 5  1.9678109 1.0075597 1.4727975 1.3495885
## 6  1.8135818 4.0246147 2.6876129 3.0063177 3.5329938
## 7  1.4751266 3.7254621 2.4798330 2.8890102 3.2247844 0.5310929
## 8  2.1457549 1.1197843 1.7627847 1.6996557 0.3528054 3.7193359 3.3768499
## 9  1.2348249 3.2471111 1.8956425 2.6317729 3.1298973 1.4929937 1.3832860
## 10 3.2235307 5.2684544 3.9937269 4.0122765 4.6052835 1.5476196 1.9175221
## 11 2.2328795 3.0916509 1.8120363 1.7885491 3.0325097 2.5759069 2.7333991
## 12 0.5404095 2.2737910 0.9335348 1.7595466 2.1579159 1.9487240 1.7174060
## 13 1.7581124 1.4296746 1.2086111 0.7212943 0.7270442 3.1047298 2.8848969
## 14 2.0639056 0.5184006 1.1030432 1.1938935 1.0003530 3.7189313 3.4730207
## 15 1.9585632 3.2263355 1.8345157 2.2792175 3.2659154 2.2639713 2.3835179
## 16 0.9492096 1.9029281 1.0351957 1.1467578 1.2700988 2.2709292 2.0035628
## 17 1.8206225 1.8304305 1.4893285 0.9235570 0.9964848 2.9588861 2.7560619
## 18 3.3571322 4.2421688 3.0169452 2.8967962 4.2013485 3.2135356 3.5232765
## 19 1.8933149 3.1515341 1.8672925 1.6129812 2.7168331 1.8984290 2.0617521
## 20 1.0109655 1.8540717 0.5130123 0.8588122 1.5905882 2.3218482 2.1637992
##           8         9        10        11        12        13        14
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9  3.3383470
## 10 4.7803380 3.0134988
## 11 3.3719878 2.0993552 3.5396879
## 12 2.3829399 0.9940760 3.4070105 1.8669425
## 13 1.0492042 2.8317257 4.0967148 2.4362375 1.8982176
## 14 1.2534032 2.9722851 4.9343875 2.5981539 1.9950542 1.1303782
## 15 3.5686049 1.3762418 3.4848713 0.9576810 1.4908738 2.7987850 2.8086905
## 16 1.4953584 2.0764335 3.3864077 2.2480068 1.2395248 0.8995979 1.6351803
```

```
## 17 1.2482197 2.8971672 3.8274144 2.5172521 2.0326133 0.4218730 1.5521435
## 18 4.5462512 2.9434471 3.8314087 1.2088960 2.9865077 3.5696791 3.7395355
## 19 3.0222731 2.0896245 2.6455367 1.1674553 1.8109290 2.0410756 2.6918487
## 20 1.8953469 1.7893429 3.5404385 1.5757798 0.9102365 1.1208814 1.4501079
##           15        16        17        18        19
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16 2.3724983
## 17 2.9179749 0.8774543
## 18 1.6912373 3.4048279 3.6071491
## 19 1.6658781 1.6880806 1.9258510 1.9692557
## 20 1.7149702 0.7694046 1.2691729 2.7753210 1.3890098
```

```
attr(distpsy, "Labels") <- pg_new$school
distpsy<- round(distpsy, digits=2)
```

Now let's create a distance matrix again but label the rows and columns by the original student ID number from the 'pg_new' dataframe.

```
pg_new_distance2<-subset(pg_new[,1:3])
(pg_new_distance2.scaled<-data.frame(sapply(pg_new_distance2, scale)))
```

```
##       control        self     motive
## 1   0.2281057  0.45967488 -0.7036200
## 2   1.6680648  0.78607125  1.0391726
## 3   0.7529506  0.01087988  0.1807822
## 4   0.2146481 -0.02991967  1.0391726
## 5   0.7529506  1.20766656  1.0391726
## 6  -1.2925988 -0.02991967 -1.5620104
## 7  -0.9696173  0.39167564 -1.5620104
## 8   0.8471535  1.54766278  1.0391726
## 9   0.1473603 -0.42431528 -1.5620104
## 10 -2.8402185 -0.02991967 -1.5620104
## 11 -0.1621636 -1.55310272  0.1807822
## 12  0.4568842 -0.02991967 -0.7036200
## 13  0.2415632  0.69087231  1.0391726
## 14  1.3316258  0.39167564  1.0391726
## 15  0.2011906 -1.49870332 -0.7036200
## 16 -0.0275880  0.69087231  0.1807822
## 17 -0.1621636  0.81327095  1.0391726
## 18 -0.6600934 -2.65469046  0.1807822
```

```
## 19 -0.9696173 -0.70991210  0.1807822
## 20  0.2415632 -0.02991967  0.1807822
```

```
distpsy <-dist(pg_new_distance2.scaled, method="euclidian", diag = FALSE, upper = FALSE)
attr(distpsy, "Labels") <- row.names(pg_new)

distpsy <- round(distpsy, digits=2)
```

Sp for example, based on the psychological measures which student is student 271 most dissimilar to , and what school do they come from ? Student 271 (School 3) is most dissimilar (3.41) to student 116 who is also from school 3.

Now let's repeat the analysis for the standardised academic measures.

```
pg_new_distance_academic<-subset(pg_new[,4:7])
(pg_new_distance_academic.scaled<-data.frame(sapply(pg_new_distance_academic, scale)))
```

```
##        english     history       maths     biology
## 1  -1.06037942 -1.85341385 -1.31590155 -1.95789911
## 2   1.57559201  1.11814381  0.96007921  1.09224688
## 3  -0.27158494 -0.12000521 -0.93122875 -0.78809880
## 4   0.25760630  0.62288420 -0.04434423  1.09224688
## 5   1.04640078 -0.06285987  1.72942481  1.56016700
## 6  -1.58957065  0.62288420 -0.23668063 -1.02205886
## 7  -0.27158494 -1.05337910  0.04113862 -1.25601892
## 8   1.57559201  0.87051401  0.24416038  0.85828682
## 9  -0.01198169  0.87051401  0.68225996  0.62432676
## 10 -1.28004370  0.12762459 -1.19836264 -0.32017868
## 11  1.04640078 -0.12000521 -0.42901703 -0.32017868
## 12 -0.75085247 -1.35815424 -0.42901703 -0.55413874
## 13 -0.80077617 -1.85341385 -1.30521619 -1.72393905
## 14  1.31598877  1.11814381  1.51571770  1.32620694
## 15  0.78679753  0.12762459  0.68225996  0.39036670
## 16  0.25760630  0.87051401  1.12035954  0.39036670
## 17 -0.01198169  0.87051401  1.30201059  0.62432676
## 18  0.25760630 -1.54863871 -1.30521619 -0.08621862
## 19 -0.53118818  0.12762459 -0.09777100  0.15640663
## 20 -1.53964695  0.62288420 -0.98465552 -0.08621862
```

```
(distaca <-dist(pg_new_distance_academic.scaled, method="euclidian", diag = FALSE, upper = FALSE))
```

```
##           1        2        3        4        5        6        7
## 2  5.5010889
## 3  2.2678863 3.4724374
## 4  4.3346727 1.7295187 2.2702846
## 5  5.4125233 1.5765960 3.7859981 2.1113142
## 6  2.9073373 4.0207039 1.6811024 2.8141330 4.2369861
## 7  1.8964298 3.8060975 1.4267566 2.9345462 3.6741568 2.1630686
## 8  4.9732107 0.7928415 2.9129566 1.3915384 1.9621102 3.7210863 3.4095357
## 9  4.3794361 1.6964208 2.3763026 0.9385633 1.9909072 2.4708591 2.7776780
## 10 2.5823844 3.9736131 1.1698801 2.4364634 4.1900200 1.3261060 2.1963422
## 11 3.3033196 2.3953268 1.4860189 1.8212634 2.8631865 2.8337006 1.9248679
## 12 1.7601694 4.0230051 1.4386328 2.7928590 3.7466074 2.2099557 1.0179722
```

```
## 13 0.3496360 5.2478326 2.0737436 4.0954764 5.1587348 2.8963171 1.7180569
## 14 5.7689428 0.6564033 3.8093620 1.9631572 1.2521403 4.1560570 4.0097601
## 15 4.1040928 1.4741589 2.2745595 1.2433483 1.6027056 2.9549629 2.3741452
## 16 4.5394482 1.5220816 2.6189630 1.3822052 1.7979924 2.7036803 2.8029826
## 17 4.6947071 1.7080917 2.8338712 1.4716041 1.7463792 2.7619312 2.9823240
## 18 2.3093887 3.9203520 1.7185865 2.7738224 3.8404905 3.1851436 1.9252089
## 19 3.1872696 2.7229571 1.3097544 1.3214122 2.7989746 1.6653925 1.8645121
## 20 3.1582687 3.8885434 1.6295264 2.3458656 4.1514703 1.1990645 2.6150576
##            8         9        10        11        12        13        14
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9  1.6634479
## 10 3.4894403 2.5668459
## 11 1.7615676 2.0562901 2.4628352
## 12 3.5815388 2.8524743 1.7703706 2.1949646
## 13 4.7048238 4.1840642 2.4771297 3.0257291 1.5439988
## 14 1.4016139 1.7355442 4.2186883 2.8457459 4.2097004 5.5291869
## 15 1.2589527 1.1156479 2.8833052 1.3669402 2.5882298 3.8555484 1.6827419
## 16 1.6503819 0.5651073 2.9660748 2.1233859 3.0457527 4.3466841 1.4878169
## 17 1.9220236 0.6197506 3.0502168 2.4474168 3.1461293 4.5115367 1.5372484
## 18 3.2987944 3.2217873 2.2891879 1.8669903 1.4282674 1.9736235 4.2643307
## 19 2.3664235 1.2840785 1.4139378 1.6991071 1.6942225 2.9984687 2.8922257
## 20 3.4882823 2.3829737 0.6427161 2.7573536 2.2526414 3.0761777 4.0800428
##           15        16        17        18        19
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16 1.0118593
## 17 1.2762292 0.4005148
## 18 2.6957554 3.4587355 3.6369608
## 19 1.5492811 1.6470157 1.7319924 2.2246003
## 20 2.9433598 2.8195145 2.8511048 2.8369685 1.4517928
```

```r
attr(distpsy, "Labels") <- pg_new$school
distaca <- round(distaca, digits=2)

pg_new_distance_academic2<-subset(pg_new[,4:7])
(pg_new_distance_academic2.scaled<-data.frame(sapply(pg_new_distance_academic2, scale)))
```

```
##        english     history       maths     biology
## 1  -1.06037942 -1.85341385 -1.31590155 -1.95789911
## 2   1.57559201  1.11814381  0.96007921  1.09224688
## 3  -0.27158494 -0.12000521 -0.93122875 -0.78809880
## 4   0.25760630  0.62288420 -0.04434423  1.09224688
## 5   1.04640078 -0.06285987  1.72942481  1.56016700
## 6  -1.58957065  0.62288420 -0.23668063 -1.02205886
## 7  -0.27158494 -1.05337910  0.04113862 -1.25601892
## 8   1.57559201  0.87051401  0.24416038  0.85828682
## 9  -0.01198169  0.87051401  0.68225996  0.62432676
## 10 -1.28004370  0.12762459 -1.19836264 -0.32017868
## 11  1.04640078 -0.12000521 -0.42901703 -0.32017868
## 12 -0.75085247 -1.35815424 -0.42901703 -0.55413874
## 13 -0.80077617 -1.85341385 -1.30521619 -1.72393905
## 14  1.31598877  1.11814381  1.51571770  1.32620694
## 15  0.78679753  0.12762459  0.68225996  0.39036670
## 16  0.25760630  0.87051401  1.12035954  0.39036670
## 17 -0.01198169  0.87051401  1.30201059  0.62432676
## 18  0.25760630 -1.54863871 -1.30521619 -0.08621862
## 19 -0.53118818  0.12762459 -0.09777100  0.15640663
## 20 -1.53964695  0.62288420 -0.98465552 -0.08621862
```

```
(distaca <-dist(pg_new_distance_academic2.scaled, method="euclidian", diag = FALSE, upper = FALSE))
```

```
##           1         2         3         4         5         6         7
## 2  5.5010889
## 3  2.2678863 3.4724374
## 4  4.3346727 1.7295187 2.2702846
## 5  5.4125233 1.5765960 3.7859981 2.1113142
## 6  2.9073373 4.0207039 1.6811024 2.8141330 4.2369861
## 7  1.8964298 3.8060975 1.4267566 2.9345462 3.6741568 2.1630686
## 8  4.9732107 0.7928415 2.9129566 1.3915384 1.9621102 3.7210863 3.4095357
## 9  4.3794361 1.6964208 2.3763026 0.9385633 1.9909072 2.4708591 2.7776780
## 10 2.5823844 3.9736131 1.1698801 2.4364634 4.1900200 1.3261060 2.1963422
## 11 3.3033196 2.3953268 1.4860189 1.8212634 2.8631865 2.8337006 1.9248679
## 12 1.7601694 4.0230051 1.4386328 2.7928590 3.7466074 2.2099557 1.0179722
## 13 0.3496360 5.2478326 2.0737436 4.0954764 5.1587348 2.8963171 1.7180569
## 14 5.7689428 0.6564033 3.8093620 1.9631572 1.2521403 4.1560570 4.0097601
## 15 4.1040928 1.4741589 2.2745595 1.2433483 1.6027056 2.9549629 2.3741452
## 16 4.5394482 1.5220816 2.6189630 1.3822052 1.7979924 2.7036803 2.8029826
## 17 4.6947071 1.7080917 2.8338712 1.4716041 1.7463792 2.7619312 2.9823240
## 18 2.3093887 3.9203520 1.7185865 2.7738224 3.8404905 3.1851436 1.9252089
## 19 3.1872696 2.7229571 1.3097544 1.3214122 2.7989746 1.6653925 1.8645121
## 20 3.1582687 3.8885434 1.6295264 2.3458656 4.1514703 1.1990645 2.6150576
##           8         9        10        11        12        13        14
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9  1.6634479
## 10 3.4894403 2.5668459
```

```
## 11 1.7615676 2.0562901 2.4628352
## 12 3.5815388 2.8524743 1.7703706 2.1949646
## 13 4.7048238 4.1840642 2.4771297 3.0257291 1.5439988
## 14 1.4016139 1.7355442 4.2186883 2.8457459 4.2097004 5.5291869
## 15 1.2589527 1.1156479 2.8833052 1.3669402 2.5882298 3.8555484 1.6827419
## 16 1.6503819 0.5651073 2.9660748 2.1233859 3.0457527 4.3466841 1.4878169
## 17 1.9220236 0.6197506 3.0502168 2.4474168 3.1461293 4.5115367 1.5372484
## 18 3.2987944 3.2217873 2.2891879 1.8669903 1.4282674 1.9736235 4.2643307
## 19 2.3664235 1.2840785 1.4139378 1.6991071 1.6942225 2.9984687 2.8922257
## 20 3.4882823 2.3829737 0.6427161 2.7573536 2.2526414 3.0761777 4.0800428
##           15        16        17        18        19
## 2
## 3
## 4
## 5
## 6
## 7
## 8
## 9
## 10
## 11
## 12
## 13
## 14
## 15
## 16 1.0118593
## 17 1.2762292 0.4005148
## 18 2.6957554 3.4587355 3.6369608
## 19 1.5492811 1.6470157 1.7319924 2.2246003
## 20 2.9433598 2.8195145 2.8511048 2.8369685 1.4517928
```

```
attr(distaca, "Labels") <- row.names(pg_new)

distaca <- round(distaca, digits=2)
```

This time,Student 271 (School 3) is most dissimilar (4.21) to student 44 who is from school 1

Mantel's test

I perfrom Mantel's test between the distance matrices for the academic and psychological measures .

```
#mantel test to see if distance matrices are correlated
library(ade4)
```

```
## Warning: package 'ade4' was built under R version 3.6.3
```

```
mantel.rtest(distaca, distpsy, nrepet = 9999)
```

```
## Monte-Carlo test
## Call: mantelnoneuclid(m1 = m1, m2 = m2, nrepet = nrepet)
##
```

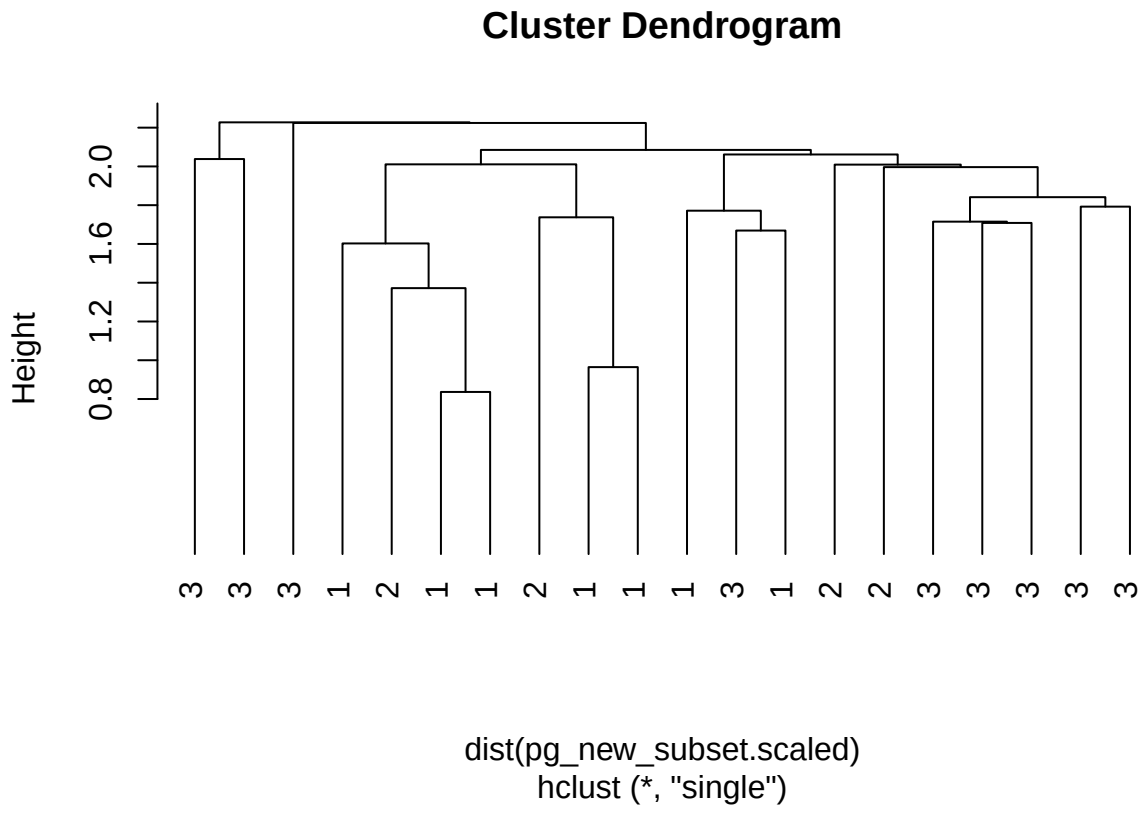
```
## Observation: 0.135583
##
## Based on 9999 replicates
## Simulated p-value: 0.0849
## Alternative hypothesis: greater
##
##       Std.Obs   Expectation      Variance
##  1.4273024265 -0.0006040362  0.0091041543
```

The Mantel test is used to determine the correlation between two matrices and whether it is significantly different to random association. This analysis indicates that the correlation between the two matrices is positive but small at 0.136. This correlation is not significantly larger than expected from a random association as p>0.05 (i.e. p=0.085).

Cluster Analysis

I perform a cluster analysis using Euclidian distances and Nearest-Neighbour linkage based on the standardised psychological and academic variables . Plot a dendrogram based on this cluster analysis and label the tips of the dendrogram branches by school.

```
pg_new_subset<-subset(pg_new[,1:7])
pg_new_subset.scaled <- scale(pg_new_subset)
pg.hc <- hclust(dist(pg_new_subset.scaled), method="single") #all vars
plot(pg.hc, hang=-1, labels=pg_new$school)
```



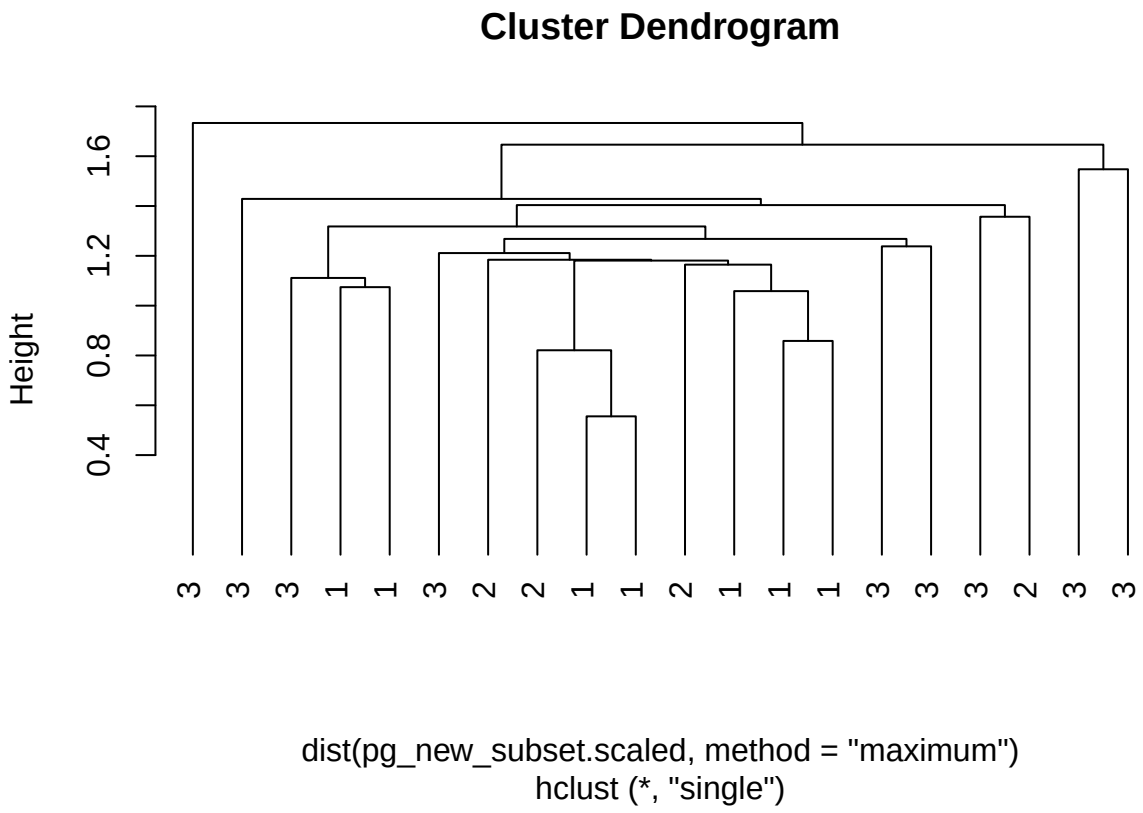
No good place to cut – branch lengths at top of tree are too short. Short branches indicate small distance between branches, so groups are not well defined.

Alternative distance measures

Minkowski - mainly useful comparing across data sets with diff number of vars. Binary - zero is failure and non-zero is success - no basis for that here dichotomy with this data Manhattan - not good for negative values which we have in the psychological data Max - takes the max difference - might help here to clarify the small branch lengths at the root Choice - maximum

Lets apply max distance method :

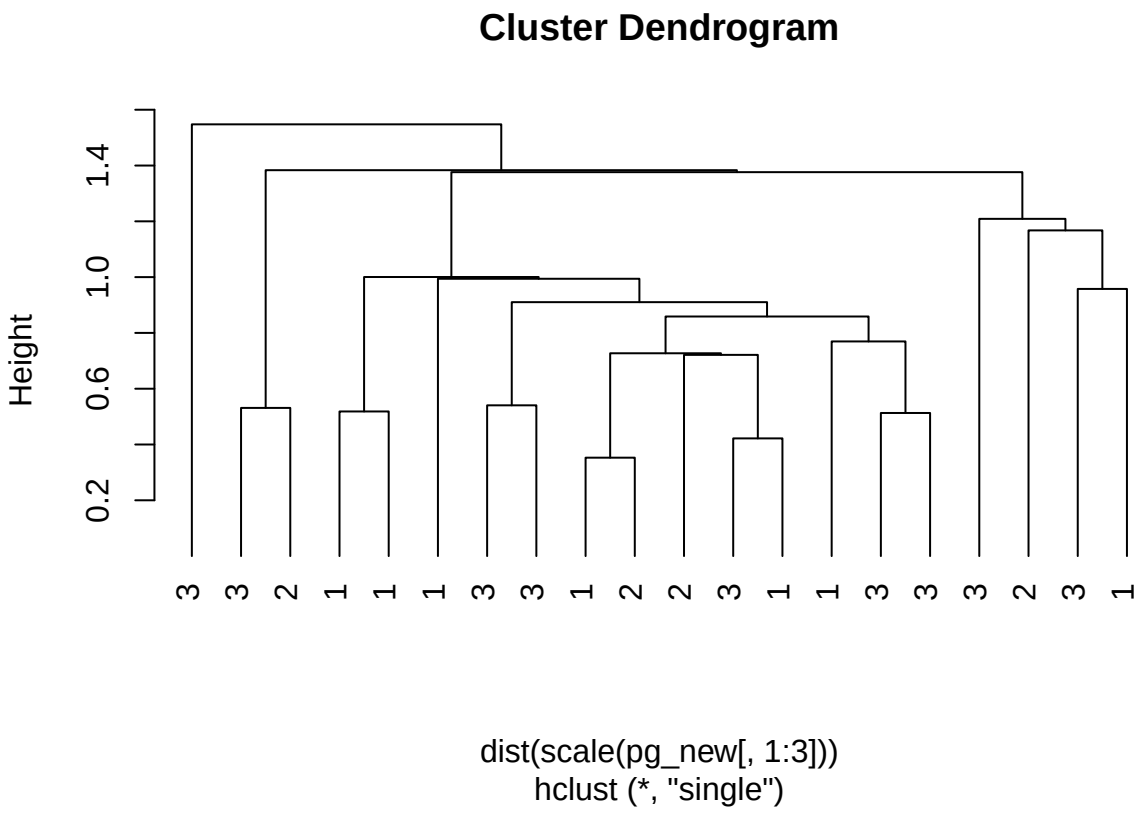
```
#standardise the data

pg.hc <- hclust(dist(pg_new_subset.scaled, method="maximum"), method="single")
plot(pg.hc, hang=-1, labels=pg_new$school)  #all vars using max distance method
```



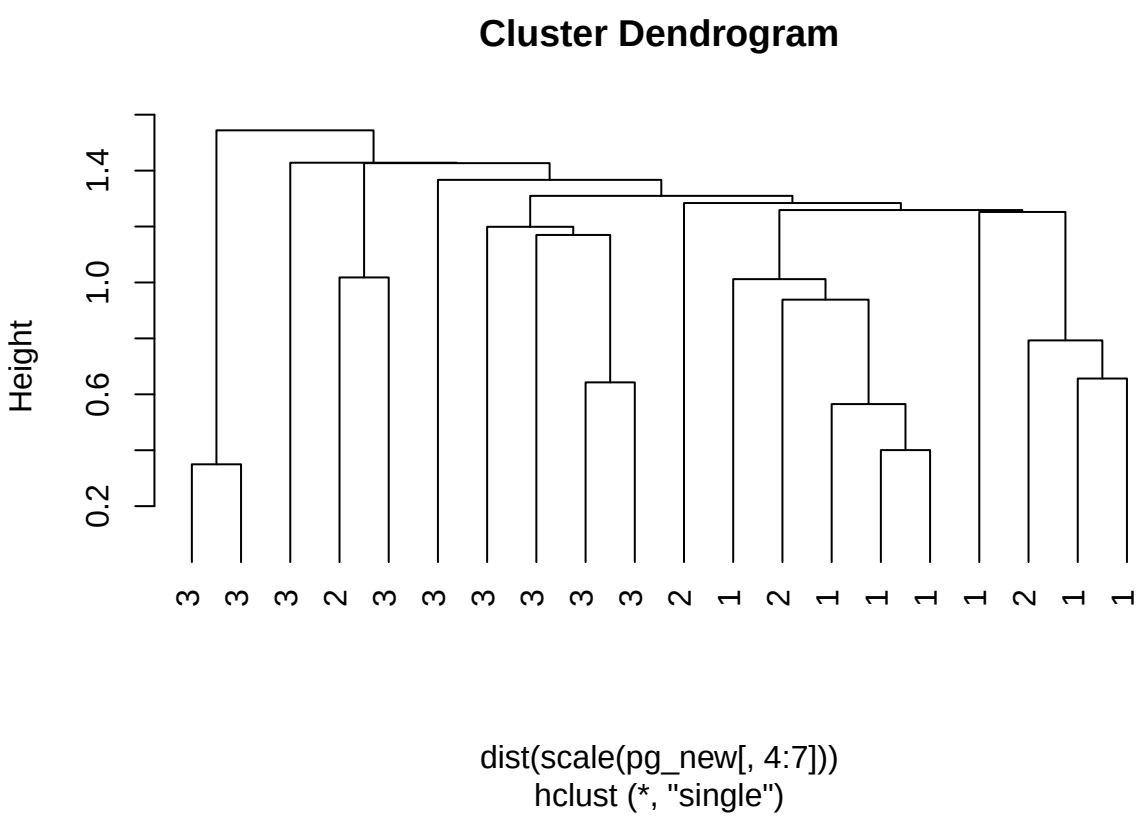
Some longer branch lengths – could cut to produce 3 or 4 groups. But the groups themselves are a mix of schools so the use of a different distance measure has not clarified relationships much more.

I Repeat the analysis for the psychological measures only and for the academic measures only.

```
pg_py.hc <- hclust(dist( scale(pg_new[,1:3]) ), method="single") #psychological vars
plot(pg_py.hc, hang=-1, labels=pg_new$school)
```




```
pg_ac.hc <- hclust(dist( scale(pg_new[,4:7]) ), method="single") #academic vars
plot(pg_ac.hc, hang=-1, labels=pg_new$school)
```



The psychological variables (first plot) do not cluster in a way that provides a clear place to cut the tree that would produce any meaningful clusters by school. The branch lengths in the academic variables dendrogram seem provide places to cut that would produce multiple clusters. School 3 would form a few clusters all only including school 3, however school 1 and 2 a re bit more mixed up. The distances between the school 3 clusters are as large as the disctances between the school 3 and school 1 and 2.

Results

Have any of these forms of analysis helped us understanding of the data?

We realized that there are a range of correlations (all positive) within and between psychological and the academic variables sets. They all have moderate correlation.Also, we realized that students who have less control and motivation are associated with students who have low academics in english , maths , history and biology. But the analysis really has not been that successful and possibly those initial correlation matrices should warn us because we had a lot of really low correlation in the Psychological matrix which was a concern. Its possible that our assumption of MVN was a mistake and maybe if we test, it won't fit a Multivariate normality.

We assumed MVN for this data and the overall redundancy indicates that while 16% of the variation in the Y set is explained by the X variables on all 3 variates and only 10% of the variation in the X set is explained by the Y variables.

Canonical Correlation has several limitations. The most important limitation is interpretability as procedures that maximize correlation do not necessarily maximize interpretation of pairs of canonical variates. Canonical solutions are mathematically elegant but difficult to interpret. In addition, the Mantel test indicates that the correlation between the two matrices is positive and we failed to reject the null hypothesis.