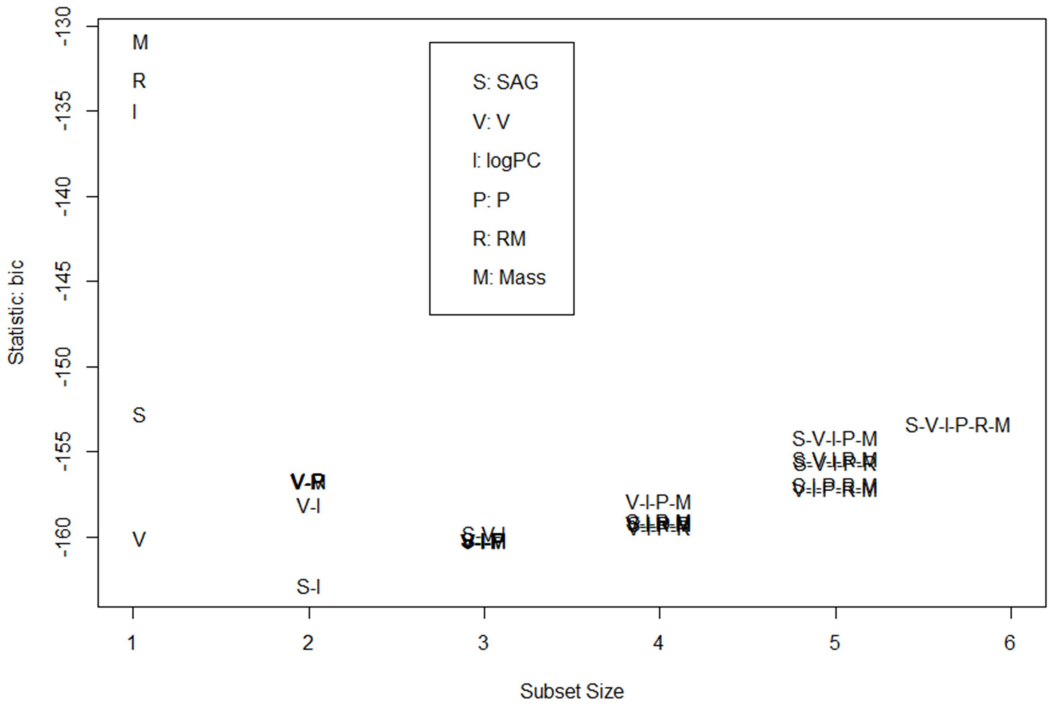
CHE384 Data to Decisions
Chris Mack, University of Texas at Austin

Homework #8 – Multicollinearity and PCA - Solution

Turn in your solution with the answers to the questions below. Also, email to me the supporting R script that you used to perform the analysis. (Please name the file using this format: HW8_yourname.R).

1. Using the alcohol data set found in DataSet4.xlsx, you wish to create a model to predict logSolubility using one or more of the calculated molecular properties of various alcohols.
   a) Search through all first-order linear models (no interactions) and find the model with the smallest BIC. Report that model.




```
Call:  lm(formula = logSolubility ~ SAG + logPC, data = alcohol)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.584765   0.721103  11.905 6.92e-15 ***
SAG         -0.027758   0.004225  -6.570 6.70e-08 ***
logPC       -1.294975   0.332417  -3.896 0.000354 ***
---
Residual standard error: 0.4648 on 41 degrees of freedom
Multiple R-squared:  0.9809,  Adjusted R-squared:   0.98
F-statistic:  1053 on 2 and 41 DF,  p-value: < 2.2e-16

> AIC(model, k=log(nobs(model))) = 69.46656
```

b) Run a correlation matrix on all of the variables. What can you conclude? Can you think of two variables that should definitely be excluded from the model? Which ones are best to eliminate?

	SAG	V	logPC	P	RM	Mass	logSolubility
SAG	1.0000000	0.9970483	0.9739759	0.9783841	0.9800329	0.9784017	-0.9868317
V	0.9970483	1.0000000	0.9862533	0.9911228	0.9920747	0.9911339	-0.9888684
logPC	0.9739759	0.9862533	1.0000000	0.9934285	0.9924252	0.9934247	-0.9802052
P	0.9783841	0.9911228	0.9934285	1.0000000	0.9998103	0.9999998	-0.9782415
RM	0.9800329	0.9920747	0.9924252	0.9998103	1.0000000	0.9998113	-0.9793722
Mass	0.9784017	0.9911339	0.9934247	0.9999998	0.9998113	1.0000000	-0.9782659
logSolubility	-0.9868317	-0.9888684	-0.9802052	-0.9782415	-0.9793722	-0.9782659	1.0000000

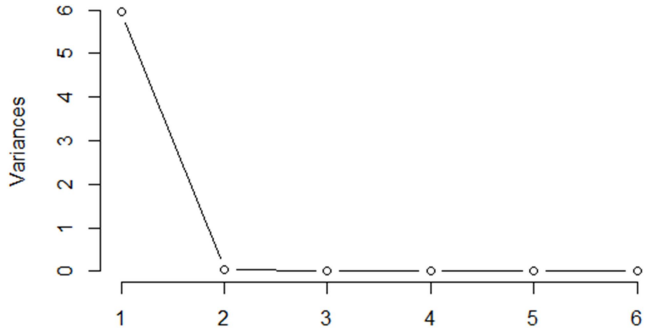
Mass, RM, and P are correlated with each other to about four nines! There is no need to keep more than one of these. I would keep Mass (the molecular weight) since the other two are derived quantities that require effort to compute.

2. Using the same data set from problem 1, perform a principle component analysis on the regressor variables. Describe what you find, and what your next actions should be.

Here are the variances explained by the six principle components and the Scree plot:

Importance of components:

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	2.4391	0.20069	0.10131	0.01309	0.01048	0.0004212
Proportion of Variance	0.9915	0.00671	0.00171	0.00003	0.00002	0.0000000
Cumulative Proportion	0.9915	0.99824	0.99995	0.99998	1.00000	1.0000000



Almost all of the variance is explained by the first principle component. The rotation matrix shows that this principle component is about the average of all the variables, and indication of very strong multicollinearity.

	PC1	PC2	PC3	PC4	PC5	PC6
SAG	0.4054016	0.7417819	-0.07216660	-0.23999603	-0.471808114	-2.626312e-04
V	0.4088248	0.3720211	0.02446446	0.42680981	0.715330295	-8.846132e-05
logPC	0.4075900	-0.3241685	-0.85096810	-0.06790216	0.005262555	2.764207e-04
P	0.4091871	-0.2776213	0.27475818	0.31319291	-0.285830228	-7.072564e-01
RM	0.4092831	-0.2288616	0.34485925	-0.74774498	0.319466920	3.695562e-04
Mass	0.4091888	-0.2771704	0.27500327	0.31370725	-0.286206124	7.069569e-01

A principle component regression, however, shows that the lowest BIC comes from the use of the first three principle components.