EXERCISE USING K-MEANS

The purpose of this exercise is to get you used to using R, or a substitute for it of your choice, to manipulate a fairly large data set and to apply a "machine learning" style of data analysis: looking for patterns in the data without starting from a probability model, and using judgment to a calibrate an index of model complexity.

The data are a subset of those used by Angrist and Krueger (1991) to study returns to schooling. It is a sample from the US census of men born 1930-1939, with data on their years of schooling, their (log) weekly earnings, their quarter of birth, and their place of birth. The file `asciiqob.txt` contains the data, one observation on each line, in the order log wage, years of schooling, year of birth, quarter of birth, and location code. There is also a file `akdata.RData` that has the data in R's data format, so that the data columns are labeled as soon as it is read in with the **`load`**`()` command. (The text file can be read in to R with the **`scan`**`()` function. but then it has to be dimensioned as a matrix and labeled.)

You are to use R's `kmeans()` function to allocate the AK data on log wage and years of schooling to groups, comparing results for 2, 3, and 4 and 5 groups (i.e. values of k). Before running `kmeans()`, scale the data so that both the wage and the schooling data have the same sample variance, i.e. the same sum of squared deviations from their sample means. The R code could be

```
akv1 <- akdata[ , 1:2]
sigakv1 <- apply(akv1, 2, sd)
## applies the  sd() command to the columns of akv1
akv1 <- akv1 %*% diag(1/sigakv1)
## normalizes so variances are the same
kout <- kmeans(akv1, 2, iter.max=100)
plot(akdata[ , 2], akdata[ , 1], col= kout$cluster)
## scatter plot with groups colored.
kout$centers %*% diag(sigakv1)
## restores original units of measurement, shows group means
```

(To see the sensitivity of the results to scaling, try this also for k=3 with the unscaled log wage and schooling data.)

Do the results suggest any substantive insight into the structure of the data, beyond the idea that income and schooling are positively related?

References

Angrist, J. D. and A. B. Krueger (1991): "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics*, 106, 979–1014.

Date: August 30, 2020.