Assignment 2

The data is taken from two sources,

https://sites.google.com/site/joseantoniocheibub/datasets/democracy-and-dictatorship-revisited as described in Cheibub, Jose Antonio, Jennifer Gandhi, and James Raymond Vreeland. 2010. "Democracy and Dictatorship Revisited." Public Choice, vol. 143, no. 2-1, pp. 67-101. DOI: 10.1007/s11127-009-9491-2.

and

http://info.worldbank.org/governance/wgi/index.aspx#home

The data are from 2008. We will use LDA to describe how countries of different "regimes" differ on a range of measures. The regime types are:

0 = Parlimentary Democracy (eg New Zealand)

1 = Semi-presidential democracy (eg France)

2 = Presidential democracy (eg USA)

3 = Civilian Dictatorship (eg Iran)

4 = Military Dictatorship (eg Cuba)

5 = Royal Dictatorship (eg Qatar)

Several governance measures are included for each country. Each of these has both an Estimated Score and a standard deviation associated with this estimate. These measures are:

VA = Voice and Accountability

PS = Political Stability, No Violence

GE = Government Effectiveness

RQ = Regulatory Quality

RL = Rule of Law

CC = Control of Corruption

In addition, the following variables are included

pacl_country: Name of the country.

bornyear: "Year the country is first identified as such." Not entirely clear what this means as New Zealand's date is 1920, which does not correspond to becoming a Dominion (1907), or becoming fully independent under the Statute of Westminister in 1931, adopted by New Zealand in 1947.

un_region_name: Region as defined by the United Nations.

tenure08: How long the current head of state had been in power as of 2008.

agereg: How long the current style of goverment has been in effect, as of end 2008. Earliest starting year is 1870, so many countries have 139 as their agereg.

Question 1: Perform the linear discriminant analysis, using numeric variables only (not region or country labels) and use leave-1-out cross validation to evaluate how many discriminant functions are useful for separating the government types. You may find it useful to add 1 to the regime variable before beginning—the LDA output will label the categories 1-6, not 0-5. Provide a pairs plot of the useful discriminant scores. Are the groups well separated? If not, which are hard to distinguish? Are any discriminant scores associated with separating particular government types?

Question 2: Compute the correlation of the useful discriminant functions with the original variables. Display the output in readable form. What are the most important variables associated with each function? Are there any variables that don't help discriminate between the government types?

Question 3: Consider the observations in "newdata.csv." What would be the posterior probability of each regime type for these observations? Display in a readable form. These are actually data from 1996. What cautions do you have about interpreting these posterior probabilities?

Question 4: Assess the normality of the LDA scores **within each regime** for your "useful" components. My suggestion for this would be to use qqplots. How do these results influence your interpretation of your results above?

Question 5: Try QDA for this dataset. You will need to exclude "royal dictatorship"—explain why. What situation is QDA designed for? Is there evidence that our data fall into this situation? How does the performance compare to LDA?