Homework 5

Knitted files in Word or PDF format should be submitted to the appropriate dropbox on My Learning Space. Please include the .Rmd file as well. When I mark your homework, the .Rmd must be able to be knitted into a report showing your R code, results of the calculations, and any necessary plots.

Download the "crimedata.csv" and "crimedata10.csv" files from My Learning Space and read the datasets into R. These files give US crime statistics by state for two time periods 10 years apart. See this link for the original data source. The following table describes the variables in these datasets.

Variable	Description	
CrimeRate	Crime rate (number of offences per million population)	Continuous
Youth	Young males (number of males aged 18-24 per 1000)	Discrete
Southern	Southern state 1 = yes, 0 = no	Binary
Education	Education time (average number of years schooling up to 25)	Discrete
ExpenditureYear	Expenditure (per capita expenditure on police)	Continuous
LabourForce	Youth labour force (males employed 18-24 per 1000)	Discrete
Males	Males (per 1000 females)	Discrete
MoreMales	More males identified per 1000 females 1 = yes, 0 = no	Binary
StateSize	State size (in hundred thousands)	Discrete
YouthUnemployment	Youth Unemployment (number of males aged 18-24 per 1000)	Discrete
MatureUnemployment	Mature Unemployment (number of males aged 35-39 per 1000)	Discrete
HighYouthUnemploy	High Youth Unemployment 1 = yes, 0 = no (high if Youth >3*Mature)	Binary
Wage	Wage (median weekly wage)	Continuous
BelowWage	Below Wage (number of families below half wage per 1000)	Discrete

1. Work through the regression tree example I provided on My Learning Space. You should see an R Markdown file and the associated data files. The example is adapted from https://rpubs.com/mammykins/reg_tree_wine.

2. Now let's build a regression tree to help predict crime rate according to some of the predictor variables found in crimedata.csv. Make sure to include expenditures on police in each state (ExpeditureYear) as well as the population size of the states (StateSize) and the number of low income families in each state (BelowWage). Also, include at least three other variables you think might be important predictors of crime rate.

3. Print a summary of your regression tree results. Which variables were most important in your tree?

4. Using the rpart.plot function, plot your regression tree. Provide a brief description in words of what the diagram is telling you.

5. According to the plot from Q4, what are the mean crime rates for each group?

6. Were any predictor variables excluded from your model? Why do you think they were excluded (i.e. how does the rpart function decide which variables to include)? This might take a little research on your part to figure out how the function works.

7. Now, let's use the model we developed with crimedata.csv to predict crime rates 10 years later. Import the crimedata10.csv file into R and use the predict function to predict crime rates given all of the predictor variables available in crimedata10.csv. See the red wine example posted on My Learning Space for help.

8. Run a Pearson correlation test using the cor function to determine how the predicted crime rate correlated with the actual crime rate listed in the crimedata10.csv file. What was the correlation coefficient from this analysis?

9. Calculate the mean absolute error of your predictions. Do you think our model was very good at predicting crime rates?

10. Using the wine example as a starting point, figure out how to generate a null distribution for comparison with your model results. In other words, if you randomly assign crime rates to states and calculate the absolute error repeatedly, what would be the mean absolute error you would attain. An annotated example is provided in the last code chunk of the red wine example.

11. What is the mean absolute error when crime rates are randomly assigned to states (i.e. your output from Q9)? How does this compare with the error from your regression tree model?

12. Is the mean absolute error from your model significantly different from random chance? Provide the p-value estimated from comparison with your null distribution.