IPSA Multi-Method: Lab 10

Due on Tenth Day

Jason Seawright

Problem 1

CART and Discovering Theory To begin, load the package rpart, which fits classification and regression trees to the data:

```
install.packages("rpart")
library(rpart)
```

Now load King and Zeng's massive dataset on possible predictors of state collapse:

```
kingzeng <- read.csv("S:/2J - Mixed Methods/kingzeng.csv")
```

The website for these data is at `https://dataverse.harvard.edu/dataset.xhtml?persistentId=hdl:1902.1/RPQIODIANR`, where you can also find a codebook.

We will use the data to analyze the combinations of causes that best predict homicide rates across countries and over time. The homicide rate is recorded in the dataset under the variable name homratet. All variables (1000+ of them) are explained in the codebook on the website for the data. To start with, predict homicides using population in cities with 25,000+ inhabitants (bnkv10), civilian vs. military government (bnkv125), energy consumption per capita (bnkv34), calories per capita per day (faocalry), and real GNP (gdf gnp) as independent variables:

```
homicide.cart1 <- rpart(homratet ~
bnkv10+bnkv125+bnkv34+faocalry+gdf_gnp, data=kingzeng)
```

Examine the results visually:

```
plot(homicide.cart1)
text(homicide.cart1)
```

We can manage the tradeoff that exists between capturing genuine complexity and over-fitting the data by pruning the tree, which means cutting off the least important parts until the results fit the data well but meet a simplicity criterion:

```
homicide.cart1.pruned <- prune(homicide.cart1, cp=0.1)
```

How do the results compare to the original tree? Explore different values of cp to see how that parameter affects results.

The data set includes a vast range of additional variables. Try running the tree using all the variables. What do you find?

Once you have a satisfactory tree, find a theoretically intriguing path through the tree. Locate a case that falls on that path, and use the qualitative evidence you can find through the internet to evaluate which, if any, of the variables on the path through the tree might be part of a causal explanation of homicide rates in that case.