POLS0012 Causal Analysis: Practical Session 1

In a famous paper titled "Islam and Authoritarianism", Steven Fish asks whether Muslim societies are less democratic.[1] To find out, he runs a series of cross-sectional regressions of countries' Freedom House scores (an indicator of the level of a country's democracy) on characteristics of the countries, including whether they are predominantly Muslim.

The paper's dataset is in the spreadsheet "fishdata.csv", which you should load using the `read.csv()` command. It contains (amongst others) the following variables:

- *FHREVERS* - Freedom House scores, a measure of democracy where higher values indicate that a country is more democratic and lower values indicate greater authoritarianism

- *MUSLIM* - =1 if a country is predominantly Muslim, 0 otherwise

- *GDP90LGN* - the country's GDP in 1990

- *GRW7598P* - the country's average annual economic growth from 1975-98, in percent

- *BRITCOL* - =1 if the country was a British colony, 0 otherwise

- *OPEC* - =1 if the country is a member of the OPEC group of oil-exporting countries, 0 otherwise

a) First, we'll practice taking subsets and summarising variables:

   i) How many countries are predominantly Muslim?

   ii) What percentage of countries are predominantly Muslim?

   iii) How many countries have GDP in 1990 of above 3.0?

   iv) How many countries are both Muslim and a former British colony?

   v) How many countries have either average economic growth from 1975-98 of above 0.6% or GDP in 1990 of above 2.5?

   vi) Create a new dataset consisting only of countries that are both Muslim and a member of OPEC

Code Hints:
- Use square brackets to denote subsets of a variable or dataset. You'll also need the `length()` function

---

[1] M. Steven Fish (2002). "Islam and Authoritarianism." *World Politics* 55 (1): 4-37

```
sum(f$MUSLIM)  # i
sum(f$MUSLIM)/length(f$MUSLIM) # ii
length(f$GDP90LGN[f$GDP90LGN>3]) # iii
length(f$MUSLIM[f$MUSLIM==1 & f$BRITCOL==1]) # iv
length(f$GDP90LGN[f$GDP90LGN>2.5 | f$GRW7598P>0.6]) # v
f.new <- f[f$MUSLIM==1 & f$OPEC==1,]  # vi
```

Answers:

i) 44

ii) 30%

iii) 88

iv) 7

v) 134

b) What is the difference in mean Freedom House score between Muslim and Non-Muslim countries? Calculate it both by hand and using a regression, verifying that your answers are identical.

```
mean(f$FHREVERS[f$MUSLIM==1]) - mean(f$FHREVERS[f$MUSLIM==0])
summary(lm(FHREVERS ~ MUSLIM, data=f))
```

Muslim countries score on average 2.20 less than non-Muslim countries.

c) Is the difference in means in (b) likely to be biased? If so, in which direction and why?

This is only a bivariate relationship, without any controls. In reality Muslim countries may be different than non-Muslim countries in many other ways that also affect their level of democracy, e.g. their level of economic development. This suggests that the measure is likely to be biased.

Some obvious omitted variables, such as economic development, are likely to be positively correlated with a country's democracy level but negatively correlated with being predominantly Muslim. The difference in means is biased downward, because we have not accounted for the fact that predominantly Muslim countries are also poorer. Other potential omitted variables are likely to be negatively correlated with the democracy level but positively correlated with being predominantly Muslim, including OPEC membership (e.g., the 'resource curse' theory in political science suggests that access to oil revenues acts as a 'curse', allowing governments to buy off citizens without introducing democracy). Again, this suggests downward bias from having failed to account for the fact that predominantly Muslim socieities are also more likely to be OPEC members.

d) Conduct a t-test for the difference in means in (b) using R's `t.test()` function. Is the difference statistically significant?

```
t.test(f$FHREVERS[f$MUSLIM==1], f$FHREVERS[f$MUSLIM==0])
```

The difference is statistically significant at any significance level, because the t-statistic is -9.63 and the p-value is extremely close to zero.

e) Conduct the t-test again, this time coding it by hand. Confirm that your answer is identical to (d)

```
d <- mean(f$FHREVERS[f$MUSLIM==1]) - mean(f$FHREVERS[f$MUSLIM==0])

se <-   sqrt(
              var(f$FHREVERS[f$MUSLIM==1])/length(f$FHREVERS[f$MUSLIM==1]) +
              var(f$FHREVERS[f$MUSLIM==0])/length(f$FHREVERS[f$MUSLIM==0])
              )

d/se
```

f) Calculate:

   i) The percentage of Muslim countries that are former British colonies
   ii) The percentage of non-Muslim countries that are former British colonies
   iii) The correlation between being a former British colony and Freedom House score

Use these results to predict what impact controlling for $BRITCOL$ will have on the estimated difference in means

Code:

```
length(f$MUSLIM[f$BRITCOL==1 & f$MUSLIM==1]) / sum(f$MUSLIM)
length(f$MUSLIM[f$BRITCOL==1 & f$MUSLIM==0]) / (length(f$MUSLIM)-sum(f$MUSLIM))

cor(f$BRITCOL,f$FHREVERS)
```

15.9% of Muslim countries are former British colonies and 23.1% of non-Muslim countries are former British colonies: The two are negatively correlated. However, the correlation between being a former British colony and the Freedom House score is virtually zero. This means that controlling for $BRITCOL$ is unlikely to have any impact on the estimated difference in means. This makes sense: the literature on colonial legacies also predicts a low correlation between democracy and colonial status, since colonialism had both negative and positive consequences for subsequent democratisation in different countries.

g) Now estimate a regression of *FHREVERS* on *MUSLIM* and *BRITCOL*. Do the results make sense?

```
summary(lm(FHREVERS ~ MUSLIM + BRITCOL , data=f))
```

As expected from the results of (f), the estimated coefficient on *MUSLIM* is virtually unchanged compared to (b).

h) Repeat (f) for *OPEC*, *GRW7598P* and *GDP90LGN*. For the latter two variables, simply calculate the correlation between each one and *MUSLIM* instead of the percentage for steps (i) and (ii)

```
length(f$MUSLIM[f$OPEC==1 & f$MUSLIM==1]) / sum(f$MUSLIM)
length(f$MUSLIM[f$OPEC==1 & f$MUSLIM==0]) / (length(f$MUSLIM)-sum(f$MUSLIM))
cor(f$OPEC,f$FHREVERS)

cor(f$GRW7598P,f$MUSLIM)
cor(f$GRW7598P,f$FHREVERS)

cor(f$GDP90LGN,f$MUSLIM)
cor(f$GDP90LGN,f$FHREVERS)
```

There is a positive correlation between being a Muslim country and being an OPEC member, and a negative correlation between being an OPEC member and democracy. On the other hand, there is a negative correlation between GDP growth/level in 1990 and being Muslim, and a positive correlation between GDP growth/level in 1990 and democracy. This means that omitting any of these variables will lead to downward bias in the estimated effect of *MUSLIM*, as explained in (c). Including them should lead to the difference becoming more positive (i.e., closer to 0).

i) Now estimate a regression of *FHREVRERS* on *MUSLIM*, *BRITCOL*, *OPEC* and *GRW7598P*. Again, do the results make sense?

```
summary(lm(FHREVERS ~ MUSLIM + BRITCOL + OPEC + GRW7598P + GDP90LGN , data=f))
```

As expected, the coefficient on *MUSLIM* is now substantially closer to zero