# Homework 1

Madi Wickett

# Question 1

**Whether ethnic minorities which exist in a society ruled by non-coethnic majority group are more likely to suffer higher risks of ethnic violence."** Assuming “higher risks of ethnic violence” is an already measured data point, regression is best suited for this question because it is trying to map if an input x (ethnic minorities ruled by non-coethnic majority societies) results in an output y (higher risks of ethnic violence). Regression works here because we are trying to model the data under the assumption that we know the data gathering process.

**“What is the likelihood that Myanmar will suffer ethnic violence next year which will leave 1000 or more dead?”** This question requires the use of statistical learning because it is trying to make a prediction about future risk. Regression cannot be used to make predictions because they will be based on how the model would handle the data, not how nature would produce the data.

**“Whether poorer countries on average suffer higher levels of ethnic violence.”** Since this question is trying to determine if there is a statistically significant difference between instances of ethnic violence in poor and rich countries, regression is sufficient to find a simple correlation between wealth and violence.

**“What region of the world is at highest risk for outbreaks of ethnic violence?”** Risk assessment is a type of question that is suited for statistical learning because you are essentially trying to predict if an event will happen sometime in the future. For the same reasons as the Myanmar question, statistical learning is better for prediction because it is trying to emulate how nature produces outcomes, not how a model produces outcomes.

# Question 2

Statistical significance does not indicate that the independent variables predict the dependent variable for a few reasons. First, when regression models are used to predict, they predict how the model would generate data, not how nature would truly generate it. Second, regression models can be overfitted to the data which will not yield accurate results when used on out-of-sample data (i.e., the future). Lastly, while all three variables may be statistically significant, one may have more predictive power than the others and the removal or inclusion of one may have little to no effect on accurate predictions - as is demonstrated by Ward et. al’s study.

# Question 3

The Curse of Dimensionality is at odds with the mainstream methodology of social science because the current thinking is to test several possible explanations for a phenomenon and see which ones are statistically significant. The Curse of Dimensionality says when you try to model so many factors, you lose precision and prediction value. Some common issues that arise because of CoD are overfitting the model and poor prediction accuracy. This creates problems for regression models like OLS and logit because we cannot use the information from our regression models to make any generalizations about the data or the future. A

statistical model estimated on one set of data might generalize poorly to new data because the model is overfitted to the training data, meaning it tried to model the training data so exactly that it did not truly pull out the important factors and won't perform well on testing data. To overcome these issues, analysts can reduce the number of dimensions their data is modeled after or cluster data.

# Question 4

## Recreating Model 1

Using a codebook available here: https://www.rdocumentation.org/packages/DirectEffects/versions/0.2/topics/civilwar (https://www.rdocumentation.org/packages/DirectEffects/versions/0.2/topics/civilwar) I decided to use onset as the dependent variable in order to properly replicate Fearon and Laitin's Model 1. In Fearon and Laitin (2003), they indicate that their first model is based on the onset of the war as the dependent variable (pg 82).

```
model1<-glm(formula = as.factor(onset) ~ warl + gdpen + lpop + lmtnest + ncontig + Oi
l + nwstate + instab + polity2 + ethfrac
            + relfrac,
            family = binomial(link = "logit"),
            data = fl)
stargazer(model1, type = "text", title = "Model 1 Replication",
          covariate.labels = c("Prior War", "GDP Per Capita", "Population (Log)", "Mo
untainous (Log)",
                               "Noncontiguous state", "Oil exporter", "New state", "I
nstability",
                               "Democracy", "Ethnic Fractionalization", "Religious Fr
actionalization"),
          dep.var.labels = "Civil War")
```

```
##
## Model 1 Replication
## ================================================
##                                Dependent variable:
##                            ------------------------
##                                      Civil War
## ------------------------------------------------
## Prior War                           -1.060***
##                                      (0.324)
##
## GDP Per Capita                      -0.421***
##                                      (0.081)
##
## Population (Log)                    0.279***
##                                      (0.072)
##
## Mountainous (Log)                    0.202**
##                                      (0.085)
##
## Noncontiguous state                   0.431
##                                      (0.279)
##
## Oil exporter                        0.887***
##                                      (0.285)
##
## New state                           1.584***
##                                      (0.339)
##
## Instability                         0.613***
##                                      (0.236)
##
## Democracy                            0.029*
##                                      (0.017)
##
## Ethnic Fractionalization              0.031
##                                      (0.371)
##
## Religious Fractionalization           0.335
##                                      (0.515)
##
## Constant                            -6.627***
##                                      (0.740)
##
## ------------------------------------------------
## Observations                          6,207
## Log Likelihood                      -469.738
## Akaike Inf. Crit.                    963.476
## ================================================
## Note:                      *p<0.1; **p<0.05; ***p<0.01
```

# Did Fearon and Laitin account for the Time Series Cross-Sectional relationships in their data by accounting for autocorrelation among the xi's?

No, they did not account for autocorrelation. Prior war is a lagged version of Civil War which denotes that a civil war is occuring, so they will obviously have very high correlation.

# Re-estimation of Model 1

There does not appear to be a difference between Fearon and Laitin's model and this re-estimation.

Note: miceadds is not supported by stargazer.

```
model1.reest<-glm.cluster(formula = as.factor(onset) ~ warl + gdpen + lpop + lmtnest
+ ncontig + Oil + nwstate + instab + polity2 + ethfrac
            + relfrac,
            family = binomial(link = "logit"),
            data = fl,
            cluster = fl$warl)
summary(model1.reest)
```

```
##                Estimate  Std. Error     z value      Pr(>|z|)
## (Intercept) -6.62650739 0.210188685 -31.5264706 3.768794e-218
## warl        -1.05988812 0.041176275 -25.7402626 4.143383e-146
## gdpen       -0.42133129 0.018591346 -22.6627644 1.044166e-113
## lpop         0.27876542 0.024247808  11.4965205  1.373411e-30
## lmtnest      0.20201459 0.031157767   6.4836031  8.955762e-11
## ncontig      0.43055495 0.033988679  12.6675988  8.940958e-37
## Oil          0.88686078 0.087045551  10.1884676  2.232555e-24
## nwstate      1.58383375 0.067539535  23.4504687 1.307430e-121
## instab       0.61349433 0.047328409  12.9624965  1.996426e-38
## polity2      0.02923881 0.003667553   7.9722938  1.557560e-15
## ethfrac      0.03097895 0.031024114   0.9985443  3.180155e-01
## relfrac      0.33540100 0.028446220  11.7907054  4.358747e-32
```