Name: Puhua Ye  NetID: Pye3

Q1:

Let A represent the event that I win after switching

Let B represent the event that I initially choose a correct door

$P(A) = P(A|B) \cdot P(B) + P(A|B^C) \cdot P(B^C)$

$P(B) = \frac{1}{3}$

$P(A|B) = 0$ Because if I initially choose a correct door, after switching, there is no chance I can win.

$P(B^C) = 1 - \frac{1}{3} = \frac{2}{3}$

$P(A|B^C) = 1$

$\therefore P(A) = P(A|B) \cdot P(B) + P(A|B^C) \cdot P(B^C)$

$= 0 \times \frac{1}{3} + 1 \times \frac{2}{3} = \frac{2}{3}$

So it is more likely to win if I switch the door.

Q2:

If they will play the $7^{th}$ game, each team wins 3 games of first 6 games. So, this is a binomial distribution.

Let $X$ be the number of games that team A won out of first 6 games.

$X \sim$ binomial $(n=6, p=0.55)$

$P(X=3) = \binom{6}{3} 0.55^3 \cdot 0.45^3 = 0.3032$

# Question 6:

## a)

This article mainly considers the Cold War period (late1940's to 1991) and the period just after the cold war (1991- 1999). The main point of the article is that civil wars are not more likely to break out because of a higher degree of ethnic or religious diversity, which is contrary to the conventional wisdom. And the article argues that we cannot predict whether civil war will break out based on where ethnic or other broad political grievances are strongest. However, the paper thinks that the main factor for deciding whether civil war will occur is "conditions that favor insurgency". If the conditions make radical and violent people easier to be lightly armed to practice guerrilla warfare from rural base areas, then civil war is more likely. The article also argues that financially, organizationally, and politically weak central government is also a major determining factor.

## b)

The paper identified 127 conflicts from a sample of 6610 country years. Moreover, for (1) Civil War, there are 6327 samples. For (2) "Ethnic" War, there are 5186 samples. For (3) Civil War, there are 6327 samples. For (4) Civil War(Plus Empires), there are 6360 samples, for (5) Civil War(COW), there are 5378 samples. The observation of samples represents different factors of a country in one year.

## c)

The independent variables include Prior war, Per capita income, log(population), log(%mountainous), Noncontiguous state, Oil exporter, New state, Instability, Democracy, Ethnic fractionalization, Religious fractionalization, Anocracy, Democracy. The dependent variable is coded "1" for country years in which a civil war began and "0" in all others.

For (1) Civil War: This is the most general consideration of civil war.
Regression equation: -0.954 * Prior war – 0.344 * Per capita income + 0.263 * log(population) + 0.219 * log(% mountainous) + 0.443 * Noncontiguous state + 0.858 * Oil exporter + 1.709 * New state + 0.618 * instability + 0.021 * democracy + 0.166 * Ethnic fractionalization + 0.285 * Religious fractionalization – 6.731

For (2) "Ethnic" war: For this one, the authors restrict attention to "ethnic wars". The dependent variable is the onset of wars that coded as "ethnic".
Regression equation: -0.849 * Prior war – 0.379 * Per capita income + 0.389 * log(population) + 0.120 * log(% mountainous) + 0.481 * Noncontiguous state + 0.809 * Oil exporter + 1.777 * New state + 0.385 * instability + 0.013 * democracy + 0.146 * Ethnic fractionalization + 1.533 * Religious fractionalization – 8.450

For(3) Civil War: for this model, the anocracy and democracy are added into the

independent variables for this model.
Regression equation: -0.916 * prior war – 0.318 * per capita income + 0.272 * log(population) + 0.199 * log(% mountainous) + 0.426 * Noncontiguous state + 0.751 * Oil exporter + 1.658 * New state + 0.513 * instability + 0.164 * Ethnic fractionalization + 0.326 * Religious fractionalization + 0.521 * Anocracy + 0.127 * Democracy – 7.019

For(4) Civil War (Plus Empires): This model omits the democracy variable and religious fractionalization. This model includes wars in the empires.

Regression equation: -0.688 * Prior war – 0.305 * Per capita income + 0.267 * log(population) + 0.192 * log(% mountainous) + 0.798 * Noncontiguous state + 0.548 * Oil exporter + 1.523 * New state + 0.548 * Instability + 0.490 * Ethnic fractionalization – 6.801

For(5) Civil War(COW): This model uses the COW data from 1945 to 1992.

Regression equation: -0.551 * Prior war – 0.309 * Per capita income + 0.223 * log(population) + 0.418 * log(% mountainous) - 0.171 * Noncontiguous state + 1.269 * Oil exporter + 1.147 * New state + 0.584 * instability + 0.119 * Ethnic fractionalization + 1.176 * Religious fractionalization + 0.597 * Anocracy + 0.219 * Democracy – -7.503

## d)

The coefficient values are weights to the independent variables which impact the onset of a civil war. A positive coefficient means if it increases, it will more likely cause a civil war and if it decreases, it will decrease the possibility of a civil war. A negative coefficient means if it increases, it will less likely cause a civil war, and if it decreases, it will more likely cause a civil war.

## e)

Positive: log(population), log(% mountainous), Noncontiguous state, Oil exporter, New state, instability, Democracy, Ethnic fractionalization, Religious fractionalization, Anocracy, Democracy

Negative: Prior War, Per capita income, Noncontiguous state for COW data.

We consider $p<0.05$ (*) as statistically significant.
For (1): Prior war, Per capita income, log(population), log(% mountainous), Oil exporter, New state, Instability
For (2): Prior war, per capita income, log(population), Oil exporter, New state, religious fractionalization
For (3): Prior war, per capita income, log(population), log(% mountainous), Oil exporter, New state, Instability, anocracy

For(4): Prior war, Per capita income, log(population), log(% mountainous), Noncontiguous state, Oil exporter, New state, Instability
For(5): Per capita income, log(population), log(% mountainous), Oil exporter, New state, Instability, religious fractionalization, anocracy

### f)

If the absolute value of weight of the independent variable is greater, it will have a greater impact on the dependent variable. So, for different 5 models, I will list 3 independent variable which has the greatest impact.
(1): New state, Prior war, Oil exporter
(2): New state, Prior war, Oil exporter
(3): New state, Prior war, Oil exporter
(4): New state, Noncontiguous state, Prior war
(5): Oil exporter, religious fractionalization, New state

## Problem 9:

### Leave-one-out cross-validation:

The leave-one-out cross validation iterates over every training sample in the dataset, and for each iteration, the parameter of the model will be updated. In this method, one sample is passed through the model at a time, and using this sample, the model will calculate the gradient to update the parameters. Therefore, for this method, if the training set contains n samples, the parameters will be updated n * number of iterations times. Moreover, it is random in this model to select every training sample. The advantage of this method is that it does not use lots of memory and it can converge faster because it updates the parameters more frequently. However, this method uses all computational resources on only one training sample at a time, which is computationally expensive.

### K-fold cross-validation:

The k-fold cross validation divides training set into multiple groups and iterates over every group. In this method, one group, which includes a few samples (normally is 5), is passed through the model at a time, and using those samples, the model will calculate the gradient according to derivative of cost function to update the parameters. The advantage of this method is that it does not use lots of memory because of the small size for each group, and it uses the computation resources on many samples instead of only one, so it is less computationally expensive than LOOCV. Moreover, this method has more stable error gradients and convergence.

### Train-test split cross-validation:

The train-test split cross validation firstly divides the whole data set in to training set and testing set and iterates over the whole training set. In this method, the whole training data set is passed through the model at a time, and using the samples of whole training data set, the model will calculate the gradient to update the parameters. The advantage of this method is that it uses the computation resources on the whole training data set, so it does not waste any computation resources. However, because the whole training data set is passed to the model, it will use lots of memory.