Nishath Fatima
STAT411

UNDERSTANDING INTERNATIONAL CONFLICT

Using statistical methods

Motivation and Research Question:

The general understanding for most social sciences is that while we are still improving on methods to predict human behavior, we can certainly use current techniques to understand it. This leads us one step closer to the general progression of difficult topics that we have little statistical perspective on. In the case of war and conflict, it is generally considered to be random, but we can certainly cultivate explanations for the impact and occurrence of war using statistical methods. Many of the methods used in our class can accomplish this feat and provide us an insightful view into the inner dimensions of war.

For this particular project, the emphasis will focus on understanding rather than predicting war. Using Principal Component Analysis and Factor Analysis, war can be understood from a dimensional perspective. Additionally, the data set created for this project is a hypothesis itself since the variables were chosen via common understanding and perspective on war. In a previous attempt to model this data set, it was discovered that the data was not linear and yielded an adjusted R-Squared of 0.229 for a simple linear regression model. Going back to the drawing board, one important aspect that was overlooked during this attempt was knowing which variables have the most impact on the number of wars by nation. Thus, a combination of both wanting to understand the data set for meaningful insight and gathering information on general vectors that can create conflict within nations, this project's major tools will be the new techniques mentioned.

Description of the Data Set:

A cultivation of the UCDP Armed Conflict[1] dataset was required to land on the final dataset created. The columns are a compilation of conflict incidents in every nation since 1948-2021 (the data set starts from 1946 but many nations had 0 conflicts until 1948 thus these years were removed). These columns were concatenated and a total sum of conflict incidents by nation was created for the new data set.

The categories were summed based on the *Intensity Level* for each instance of conflict. Note that this means some nations will get a double count if the conflict that ensued resulted in over 1000 battle-related deaths.

[1] https://ucdp.uu.se/downloads/index.html#armedconflict

UCPD Dataset Variables:

Year: The year of conflict, ranging from 1948 - 2020

Location: A list of the country/territory where the conflict took place. This resulted in every instance of conflict being logged, for example, the war between Pakistan and India had over 45 conflicts, thus the data set would have 45 entries of Pakistan or India (depending on where the conflict had taken place)

Side_a: The primary party in a conflict, always a government regardless of whether the opposer is a government or individual group.

Side_b: The opposing actor in a conflict, can be an opposing nation or a military opposition group.

Incompatibility: The reason for incompatibility. 1: Incompatibility about government, 2: Incompatibility about territory, 3: Incompatibility about both

Intensity Level: The intensity level of the conflict. 1: 25-999 battle related deaths, 2: 1000+ battle related deaths

Region: 1: Europe, 2: Middle East, 3: Asia, 4: Africa, 5: Americas

After concatenation, the dataset used for the project incorporated 5 numerical variables to be tested against the incident count. Each nation's average Population, Land Size, Exports, Labor Force and Military Spend were attached to their respective country for the resulting data to be used.

Analysis and Results:

The most important preliminary factor to note is that during exploratory data analysis, several outliers for each variable were uncovered. This had led to the extremely low R-squared provided in the motivation section. When removing outliers from all variables, the R-Squared increased to 0.38 but had no robustness and is undoubtedly biased. Kernel Regression fitting was able to visually show that this data is certainly not linear, making it difficult to model without non-parametric methods. These analyses can also help to understand underlying relationships with the influential outliers included, which will be extremely useful.

The first, second, and third PC account for a total of 96% of variance. At an individual level, the first component accounts for 69%, the second for 17%, and the third for 10%. As discussed in lecture, one of the major use cases for Principal Component Analysis is data reduction, where " *p components are required to reproduce the total system variability, much of this variability can be accounted for by a small number k of the principal components. : these k principal components can then replace the initial p variable.".* It is important to note that though there are not several variables within this dataset at the moment, additional could be added and this use case could be particularly successful for an attempt to predict war with machine learning techniques. For dimension reduction in this case, a scree plot was used to finally assess the optimal number of components to keep for future models- three.

A correlation coefficient matrix shows that all variables have a relatively even effect on the first principal component. The next two major components, however, begin to show more impactful values. With Population, Labor Force, and Military Spend having the highest coefficients for the second component, and with Land Size being the largest by far amongst all in the third, the interpretation for the dimensions should provide us with more information. This is another significant use case for principal component analysis: "*PCA reveals relationships that were not previously suspected and allows interpretations.".* Additionally, these new relationships can bring research questions that surface from the extended knowledge on the domain. Secondly, we can further understand the relationship we have mapped out with Principal Component Analysis via Factor Analysis.

The factor loadings portray a much more straightforward story. The highest loadings for the first column are Population and Labor Force, the second for Military Spend and Exports and the third for Land Size. By the third factor 96.8% of variance is explained with the PCA method. Though this method was assumed to be best, it was the varimax rotation model that provided the most accuracy and information by accounting for 98% of variance. Despite the completely different values within the factor loadings, it seems that similar patterns come up in this model. In comparison to PCA, the interpretability within Factor Analysis promises to be more fruitful.

Interpretation and Conclusion:

The UCDP dataset provided a general scope into the patterns that can be shown for a subject such as war. 55% of battles are over territory, and 75% are facilitated by rebel groups within their population. It makes sense in this regard that population always loads as a very high correlation coefficient within the first principal component explaining a large amount of the dataset. The population of a country is difficult to manage, and when opinions are generated within this population a situation can easily escalate. This can also be due to the relationship with labor force, which could have a more practical relationship with population when showing up within the same component, as population will have a direct proportional relationship to its labor force.

The first component in principal component analysis may be considered a “people component” as it had no noticeably high coefficients for any of the variables. In comparison, the factor analysis showed Population and Labor Force as the highest in the first set. The above understanding shows that the factor analysis has likely deciphered this relationship more accurately from a domain perspective.

The second component for PCA can be interpreted as the Economic Component, as its coefficients were highest among Population, Labor Force, and Military Spend. This loading is once again more clear in the Factor Analysis, with only Labor Force and Military Spend being included as solely monetary values. Note that for both PCA and Factor Analysis, Military Spend has a negative correlation coefficient. Meaning low Military Spend and High Labor are associated with this component (in the case of PCA, population is present as well).

For the third component in both PCA and Factor Analysis, only Land Size is present. Given that 55% of wars are over territory, and 72% of total wars are within Asia and Africa, this could be a component that has significant effect but not necessarily against the entire data set and its variables. It was important to include this factor and understand that though we see an impact from Land Size in this analysis, war on land is very difficult to assess from a statistical standpoint as the reasons behind a territory war can go beyond how large the existing territory of a nation is. Additionally, the ratio to population and land size could be a more meaningful indicator to assess.

Overall, the above analyses were able to provide discernible information on the relationship between random variables that were mapped without cause in the case of war. It is clear that the insightful and interesting story both PCA and Factor Analysis were able to tell helps understand the way conflict can be asserted within the variables assigned but also the usefulness and applicability of the techniques themselves. While mapping out the intuitive relationship of the coupled variables that could be difficult to detect in a bigger data set, using PCA and Factor analysis, the ideal next step would be to complete a full predictive model and provide even more valuable information for this topic, which is certainly important to address.