Predicting Violence in Militarized Interstate Disputes

Cleary, Koruna & Yong

5/4/2019

Brief Introduction

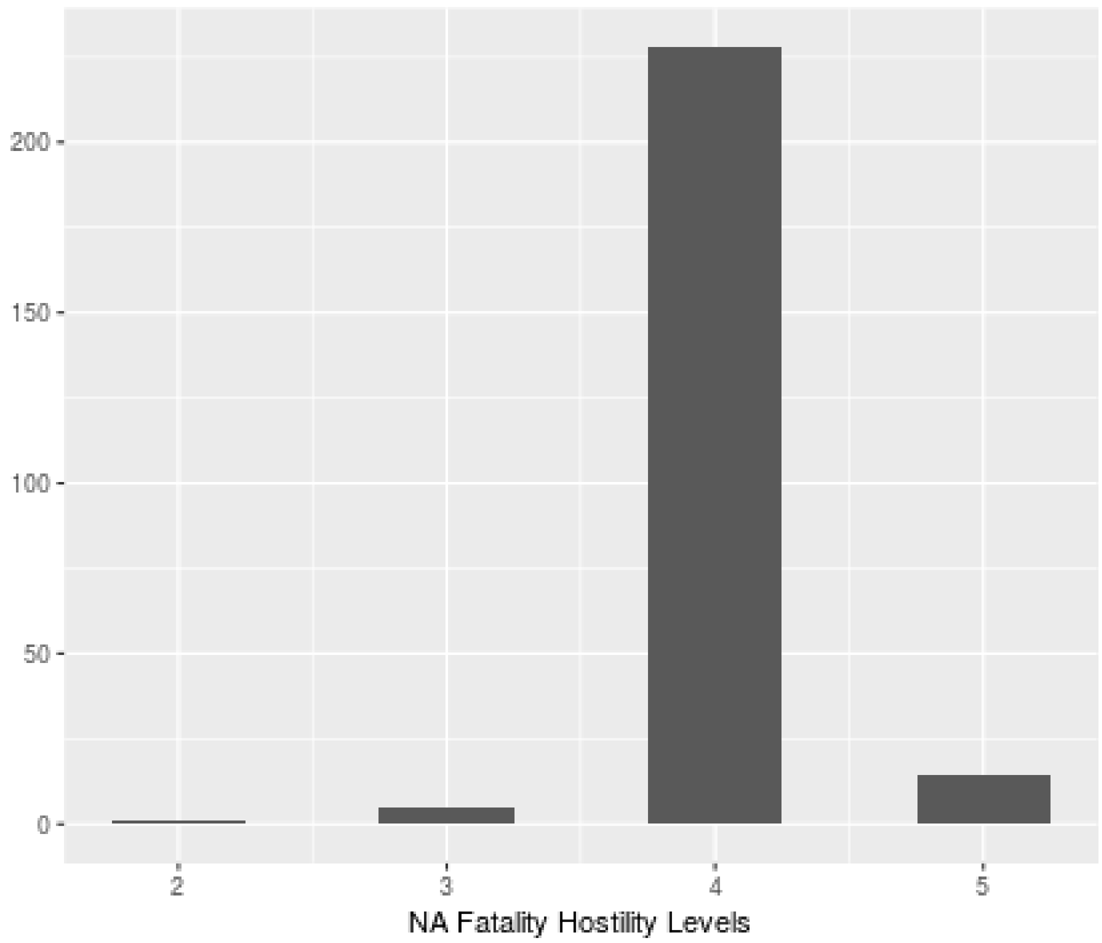
Zach Cleary's Work

Militirized interstate disputes (MIDs) are tense moments in international diplomacy. The formal definition is "Militarized interstate disputes are united historical cases of conflict in which the threat, display or use of military force short of war by one member state is explicitly directed towards the government, official representatives, official forces, property, or territory of another state. Disputes are composed of incidents that range in intensity from threats to use force to actual combat short of war" (Jones et al. 1996: 163). Whenever a MID occurs there is a risk of violence and deaths. Our objective is to attempt to predict if a given incident will results in fatalities. While the vast majority of MIDs do not result in fatalities those that do can have signifigant impacts. To do so we have drawn upon the Correlates of War extensive MID data. Said data is split into two data-sets. The A data-set is organized by incident and the B by actor.

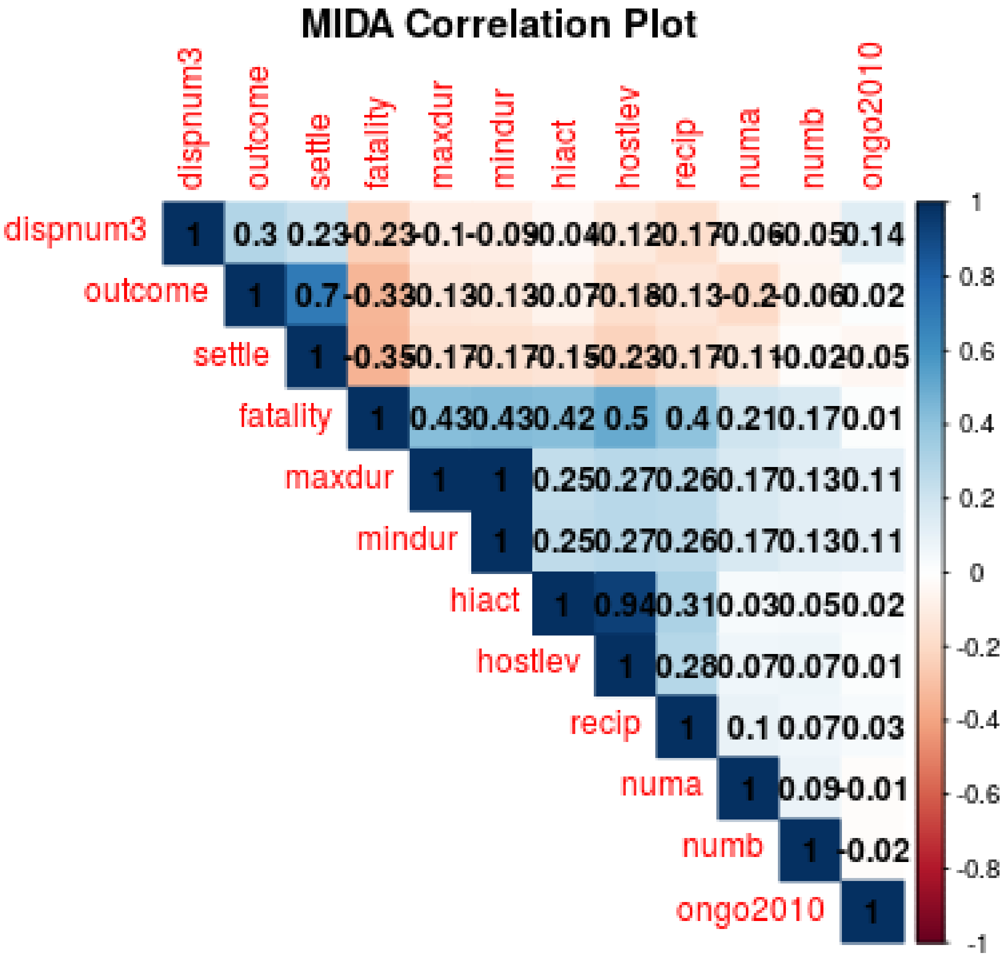
Data Exploration

David Koruna

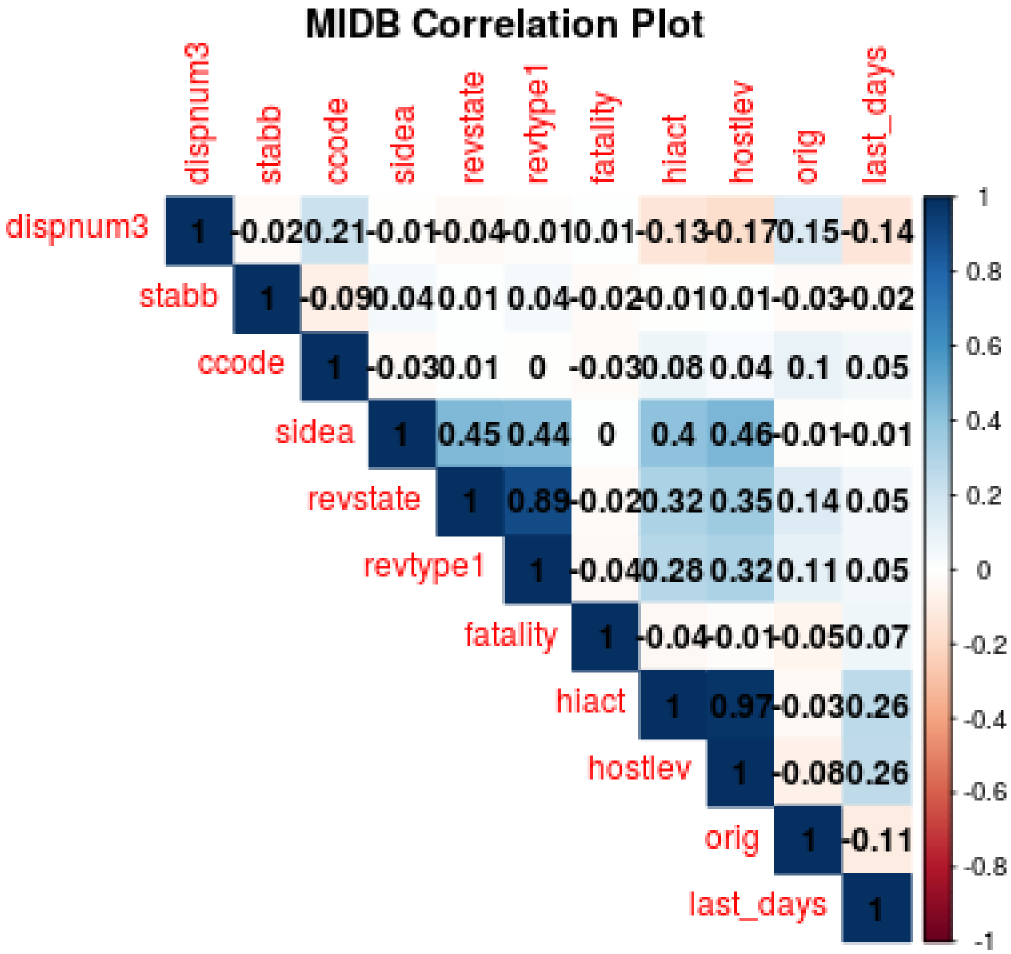
For this particular project, we are taking a look at "fatality" as the response variable for both the MIDA and MIDB dataset. For the "fatality" variable, we noticed that there were about 248 missing records out of the 2292 observations. I wanted to take a look at why that may be the case. Upon examination of the Missing NA values, I was able to find that in most cases were number of fatalities were missing, hostlity levels were at a 4 or higher. This indicated that a hostility level associated with "use of force" or higher was causing fatality reports to go missing. This can be visualized easily using a histogram:



We also wanted to take a look at the correlation between different variables in the datasets. Starting with the MIDA set:



We see high levels of correlation between “outcome” - “settle,” “maxdur” - “mindur,” and “hiact” - “hostlevel.” These correlations all make sense. The outcome of the dispute increases as settlement increases. Min duration and Max duration would scale together, as estimated duration of the dispute is computed in an interval. It also makes sense that highest action taken increases with hostility level, as these variables are directly related with each other.



	dispnum3	stabb	ccode	sidea	revstate	revtype1	fatality	hiact	hostlev	orig	last_days
dispnum3	1	-0.02	0.21	-0.01	-0.04	-0.01	0.01	-0.13	-0.17	0.15	-0.14
stabb		1	-0.09	0.04	0.01	0.04	-0.02	-0.01	0.01	-0.03	-0.02
ccode			1	-0.03	0.01	0	-0.03	0.08	0.04	0.1	0.05
sidea				1	0.45	0.44	0	0.4	0.46	-0.01	-0.01
revstate					1	0.89	-0.02	0.32	0.35	0.14	0.05
revtype1						1	-0.04	0.28	0.32	0.11	0.05
fatality							1	-0.04	-0.01	-0.05	0.07
hiact								1	0.97	-0.03	0.26
hostlev									1	-0.08	0.26
orig										1	-0.11
last_days											1

Figure 1: MIDB Correlation Table

Looking at the MIDB dataset:

Here, we are seeing high amounts of correlation between “revstate” - “revtype1” and “hiact” - “hostlev.” Again, these high correlations make sense and can be explained by the dataset. Whether or not a state is a revisionist state should cause a direct increase in revision type, because if “revstate” is 0 then “revtype1” is automatically set to 0 as well, as there’s no type of revision.

MIDA Models

Zach Cleary's Work

Description of MIDA and Goal

MIDA is the portion of the MID data organized by actor. The primary objective of this section is to fit a relatively simple but effective logistic regression for predicting if there will be fatalities. Several models are fitted and then compared in the concluding portion of this segment. The full summary statistics are presented in an appendix at the end of the report. However, a few variables require explation before proceeding.

- outcome: A catagorical variable that records how the conflict ended.
- settle: A catagorical variable that records manner in which the final agreement was made. Was the final settlment negotiated or imposed?
- hiact & hostlev: Strongly related ordinal variables that measure the level of escalation reached.
- deaths: The boolean objective variable. It is true if there were deaths.

Since hiact and hostlev are arguably too potent a more limited model is fitted without them.

Data Cleaning

	x
dispnum4	2030.00
stday	262.00
endday	245.00
outcome	15.00
settle	20.00
fatality	248.00
fatalpre	629.00
link3	3.00
deaths	248.00

Table 1: Counts of NA entries

Before fitting any models the data required a second round of cleaning utilizing data.table. This mostly consisted of the following actions:

- Removing purely administrative variables such as the MIDs internal ID number.
- Dropping the fatalpre and fatality variables as they were redundant given the objective variable.
- Dropping variables such as endday that were deemed uninteresting/unnecessary. The end month might give insight into fighting seasons but days have no such value.
- Omitting all MIDs with missing values. Since this step is performed last the 2030 missing dispnum4 entries are not a problem.

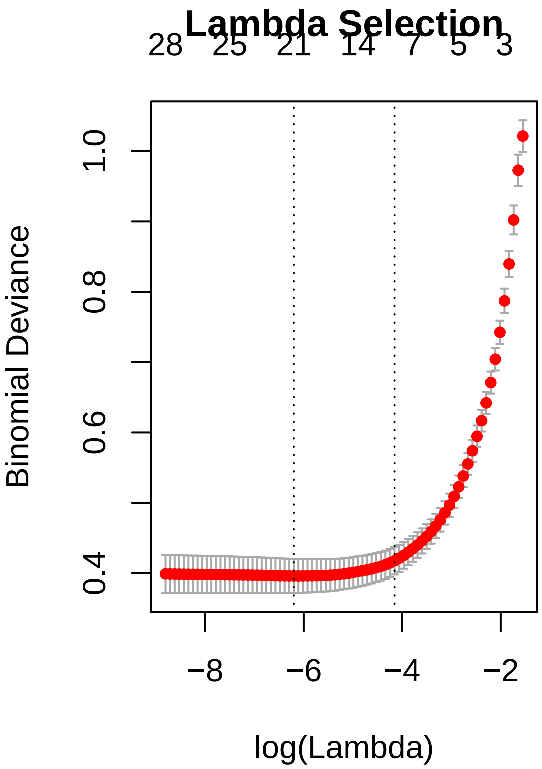
This last action was necessary for the modeling techniques I used but did result in a ~12% reduction in entries. More problematically it probably biased the resulting models as prior analysis indicates that violent incidents are more likely to having missing information.

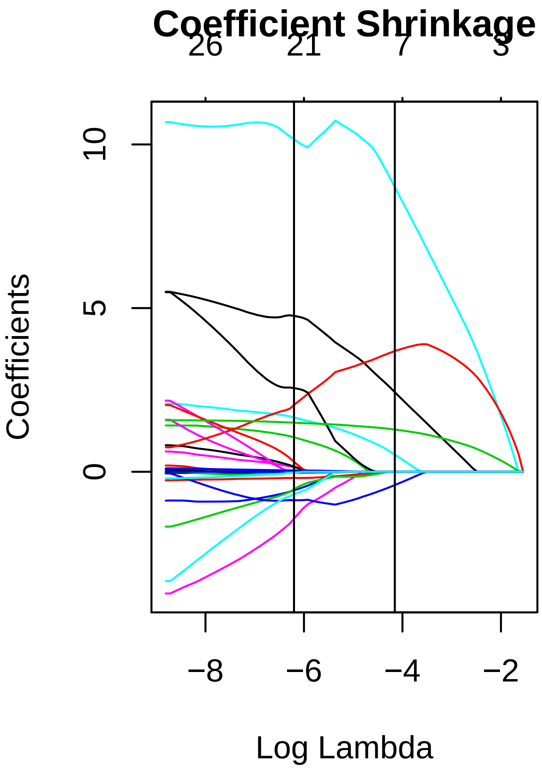
Baseline Models

Two models were created for comparison purposes. One is a simple intercept only model. The second is the product of automated step-wise selection based on the AIC starting from the intercept only model. The step-wise selection process throws out numerous warnings stating that "fitted probabilities numerically 0 or 1 occurred." This is expected behavior. The ordinal variables (hiact,hostlev) have not been converted to numeric approximations. This means that the glm function attempts to use k-1 polynomial contrasts. Unfortunately, there is not sufficient data to fully support this and thus the warnings. If the intention was to proceed with a glm() based model steps would be taken to rectify the situation. However, in this case it is intentionally left to show the improvement of the alternative method.

LASSO Models

In order to simultaneously improve results and conduct variable selection the least absolute shrinkage and selection operator (LASSO) technique was utilized in addition to the logistic regression. This was accomplished via the cv.glmnet() function of the glmnet package. While a complete explanation of the LASSO technique is beyond the scope of this project there are a few elements that require exploration. I utilized cv.glmnet()'s built in parallel comparability as the function can be computationally expensive. The cv.glmnet function fits #folds+1 LASSO logistic regressions. The first fit is used to produce a sequence of lambda values (the tuning parameter) and the rest are traditional cv fits to evaluate the model. The performance of the various lambda values can be seen below. I have chosen to select not the lambda that minimizes the mean cross-validated error but "lambda.1se" (0.016 for the main LASSO model) which is the lambda within one standard error that gives the most regularized model. This is due to desire to use LASSO as a variable selection technique. The second plot below shows how the lambda values influence the coefficients. From right to left the two lines represent the lambda which minimizes error and the lambda.1se. It is apparent that using lambda.1se results in a significantly simpler model.





Model Comparisons

Table two above shows the summarizes key measures about the performance of all the models while table three below displays their coefficients. As can be seen the step-wise model is by far the most complicated. This is why the LASSO model is preferred even though it performs marginally worse. The deviance values are calculated from 10-fold cross validation. It is clear that the key performance metric is sensitivity. Most MIDs

	Model	Accuracy	Acc p-Val	Sensitivity	Specificity	Deviance	AIC
1	Intercept	0.79	0.51	0.00	1.00	0.17	2071.74
2	Stepwise	0.93	0.00	0.81	0.96	0.06	811.88
3	Lasso	0.93	0.00	0.79	0.96	0.42	851.03
4	Reduced Lasso	0.85	0.00	0.45	0.96	0.54	1089.96

Table 2: Overview of results

do not end in deaths but correctly predicting those rare instances are crucial. Thus it is unacceptable to use a model such as the intercept with relatively high accuracy but abysmal sensitivity. It is also noteworthy that the reduced LASSO model is still relatively effective predictor even though it lacks to potent retrospective variables hiact and hostlev.

	Intercept	Stepwise	Lasso	Reduced Lasso
(Intercept)	-1.34	-20.52	-3.76	-11.61
hiact.L		39.87	8.70	
hiact.Q		12.78		
hiact.C		3.65		
hiact^4		12.74		
hiact^5		9.09		
hiact^6		-5.61		
hiact^7		-7.58		
hiact^8		-1.11		
hiact^9		7.22		
hiact^10		16.96		
hiact^11		19.41	2.43	
hiact^12		10.95	3.69	
hiact^13		3.75		
hiact^14		1.95		
hiact^15		0.71	-0.41	
hiact^16		-9.34		
hiact^17		-3.75		
hiact^18		12.06		
recipTRUE		1.61	1.29	2.44
mindur		0.02	0.00	
maxdur		-0.02		0.00
settleImposed		0.56		
settleNone		-0.26		
settleUnclear		1.41		
styear		0.00		
numa		0.15		
outcomeB victory			0.52	0.56
endyear				0.00
outcomeReleased				-1.49
outcomeJoins ongoing war				-2.03
hostlev.L				5.88

Table 3: Coefficients from all models

Modeling for B

Yong's Work

Data Cleaning

Used Dplyr,tidyr,data.table libraries to manipulate the data more easily and made the cleaning procedure faster.

- Dropped dispnum4,revtype2 because they contain too many missing values.
- Dropped fatalpre because it contians missing values and highly correlated to fatality.
- Dropped the last 4 variables which is unrelated.
- Calculated the lasting days of the dispute based on the startdate and enddate and create a new variable called last_days, if one of the day of a date is missing, the first day of the month will be applied.

- Dropped the missing rows where the fatality is missing and transfer fatality into a binary variable where 0 means no death while 1 means there is death.

After the whole cleaning process, there are **8** variables left in the data. The data dimension is **4459 X 8** .All the variables we used in this data set are categorical instead of last_days which is numeric. However, even though they are categorical, they came in consecutive integers starting from 0. We wonder if there is much difference with our logistic model if converting these 7 special categorical variables into factor data type in R.

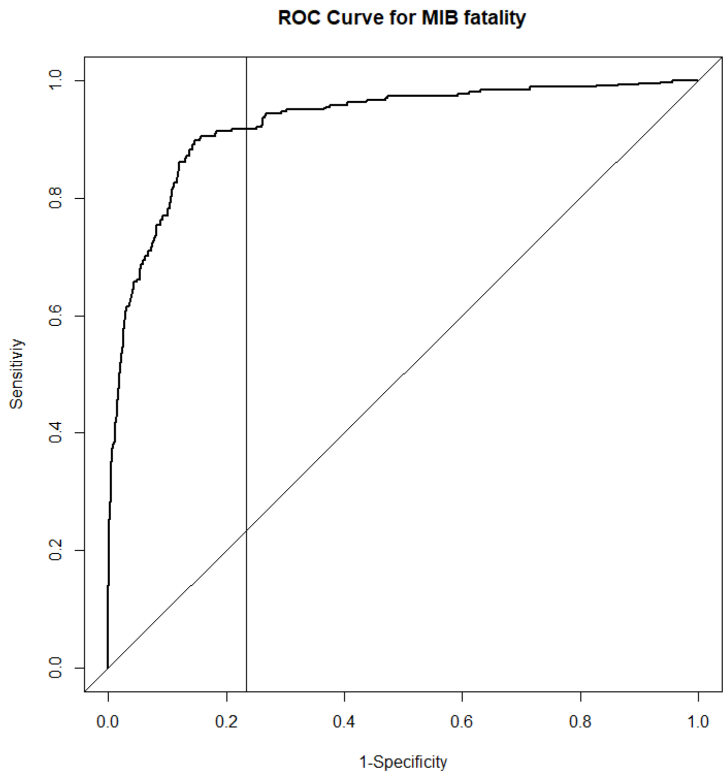
Model and Result

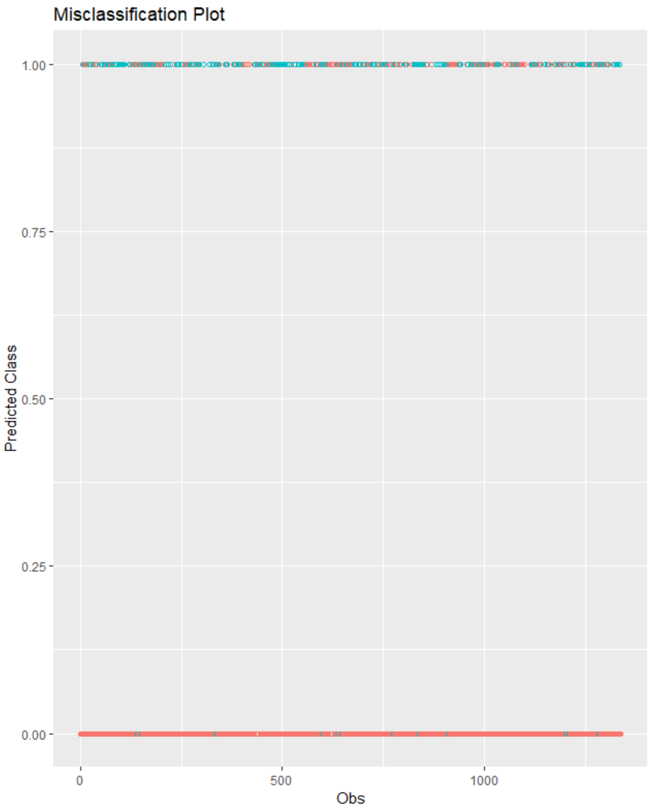
We randomly split the data into a training set and a testing test. 70% of the data randomly goes into the training set, the rest 30% goes into the testing set.

Logistic model

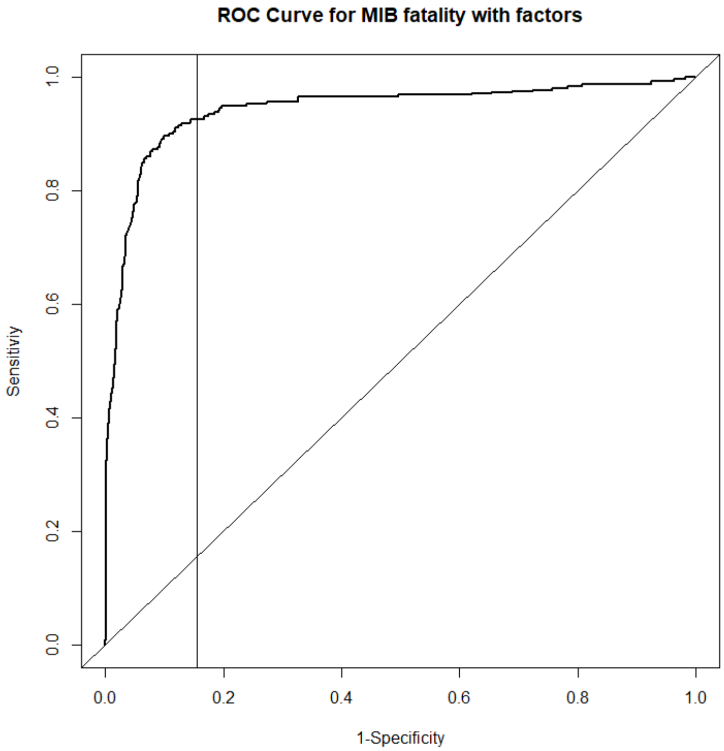
For logistic regression, we used the training set to fit the model and the test set to predict the fatality level. The "fatality" is the respond, "sidea", "revstate","revtype1","fatality" ,"hiact","hostlev","orig" and "last_days" are the predictors. After fitting the model, we calculated the AUC(Area under the curve), log-loss and misclassification rate for each model.(for each split and each model) Also, we employed ggplot2 to plot ROC curve and misclassification plot for the fitted value.

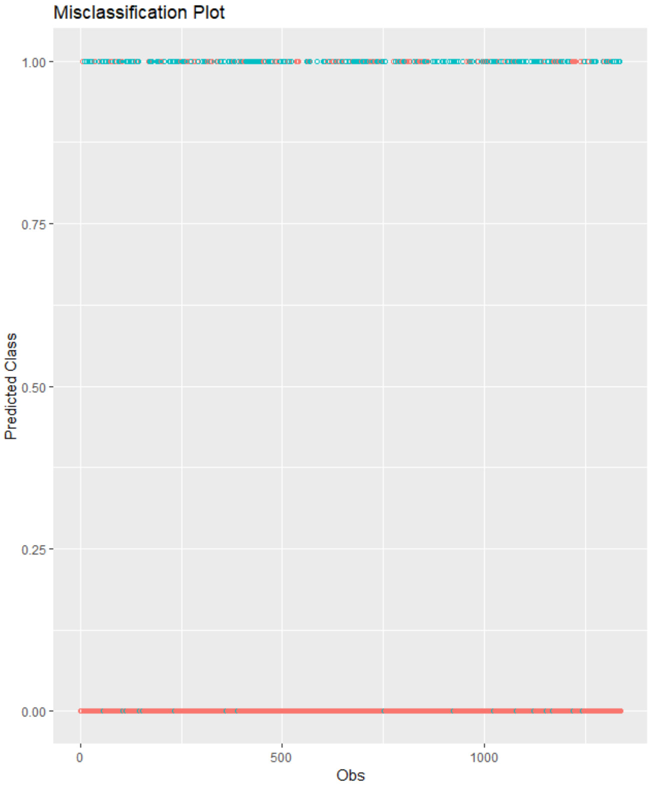
Logistic model with numeric predictors





Logistic model with factor predictors





From the ROC curve and the misclassification plot, we can see that the AUC for the two models are both good but apperantly the one with the factors are better with more curvy, meaning area under curve will be higher. As for the misclassification plot, we can easily see generally for the 0 class, there are less misclassification rate is less than class one, for the class one, the model with the factors has more way less red than the model with numeric type.

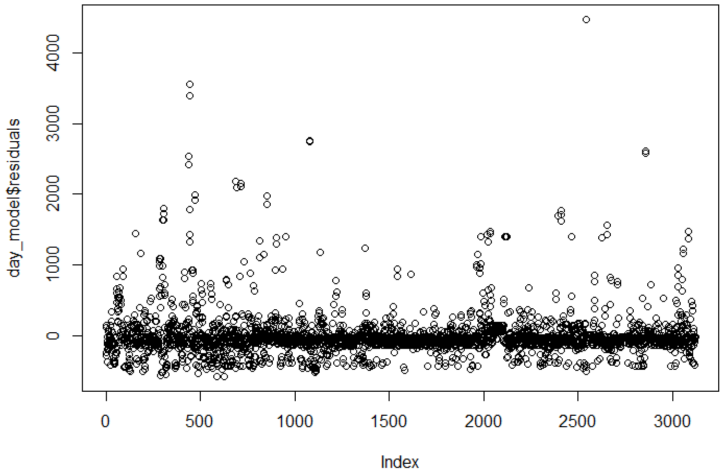
Result

Log-loss and AUC measure the performance of a classification model the lower log-loss is, the higher AUC is, the better the classifier is Apparently, the model with the factor predictors has higher AUC, lower log-loss and lower misclassification rate. What it tells us is that do transfer your categorical variables into factors even though they look very numeric.

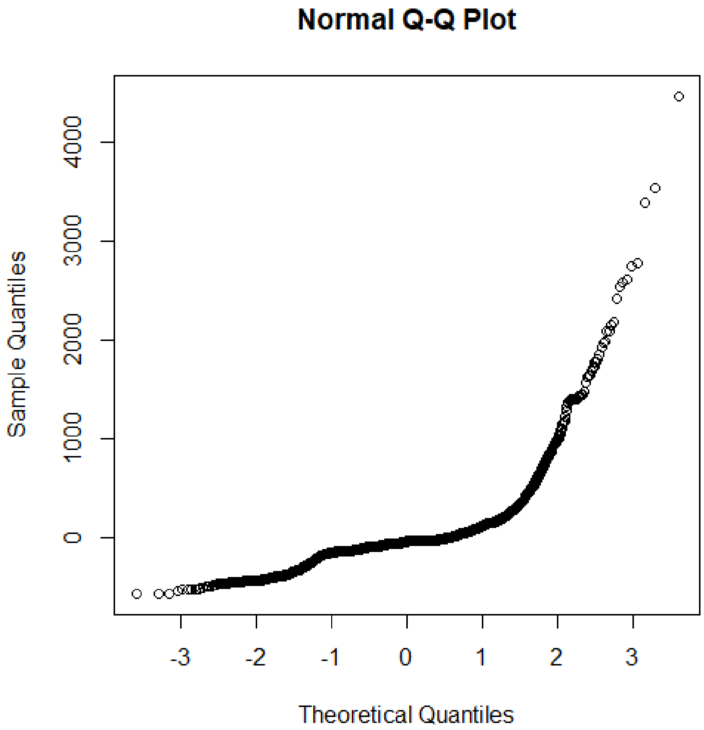
Model	AUC	Log-loss	Misclassification rate
Logi numeric	0.9249	0.2621	0.1399
Logi factor	0.9421	0.2185	0.0979

Linear regression model to predict lasting days of a MID

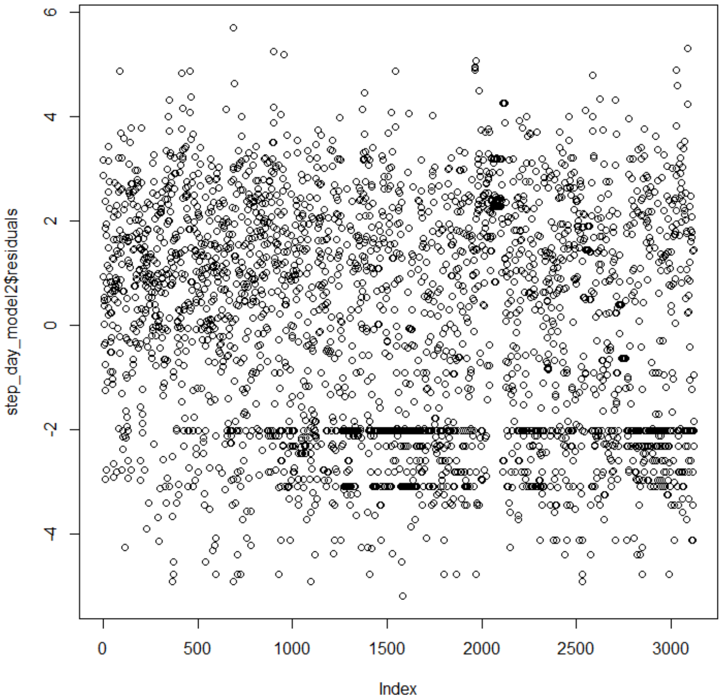
The lasting days of MIDs range from 1 to 4903, it would be very interesting to find the relationship between the lasting days and other predictors.
* We fitted a linear regression for lasting days. Then we used stepwise selection to perform model selection. We found that the residuals of the selected model with the lowest AIC turned out not following the normal distribution.
* Therefore, we transformed the respond into log-response and refitted the model.



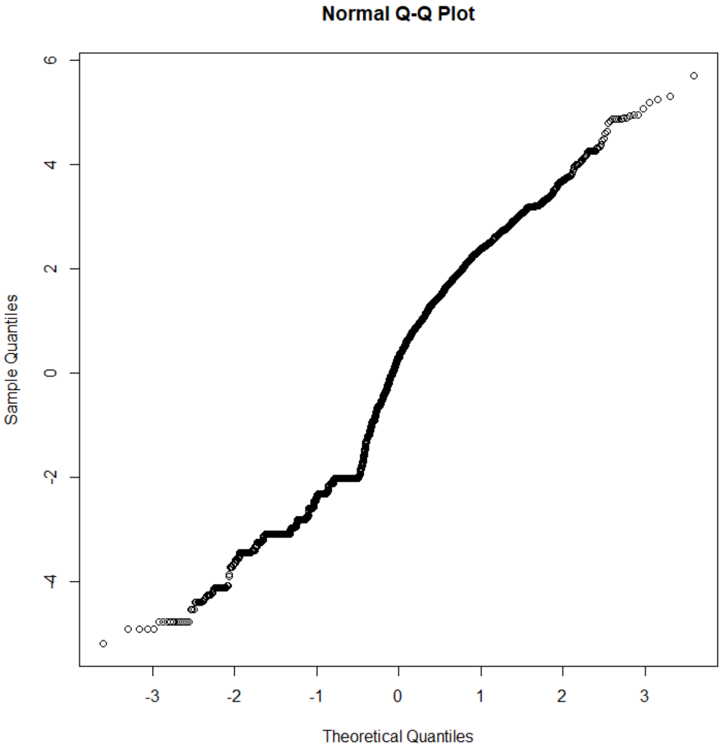
There are some outliers but overall the residuals are constant variance.



The trend is far from a straight line, therefore the residuals are not normally distributed.



After transformation, the residuals plot shows that the residuals are homogeneous with points evenly spreading out below and above 0.



Also the QQplot shows that the new residuals closely follow the normal distribution. Then we predicted the respond for the test data set and calculate for the RMSE(Root Mean Square Error) of 0.05467609 which indicates that the model is a good fit.

Citation

Palmer, Glenn, Vito D'Orazio, Michael R. Kenwick, and Roseanne W. McManus. Forthcoming. "Updating the Militarized Interstate Dispute Data: A Response to Gibler, Miller, and Little." International Studies Quarterly.

MIDA Appendix: Complete Summary

Data Frame Summary

NNA__Data

Dimensions: 2024 x 15
Duplicates: 51

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	stmon [integer]	Mean (sd) : 6.2 (3.4) min < med < max: 1 < 6 < 12 IQR (CV) : 6 (0.5)	12 distinct values	2024 (100%)	0 (0%)
2	styear [integer]	Mean (sd) : 1958.5 (43) min < med < max: 1816 < 1969 < 2010 IQR (CV) : 53 (0)	184 distinct values	2024 (100%)	0 (0%)
3	endmon [integer]	Mean (sd) : 6.5 (3.4) min < med < max: 1 < 7 < 12 IQR (CV) : 7 (0.5)	12 distinct values	2024 (100%)	0 (0%)
4	endyear [integer]	Mean (sd) : 1958.8 (43) min < med < max: 1816 < 1970 < 2010 IQR (CV) : 52 (0)	186 distinct values	2024 (100%)	0 (0%)
5	outcome [factor]	1. A victory 2. B victory 3. A yield 4. B yield 5. Stalemate 6. Compromise 7. Released 8. Unclear 9. Joins ongoing war	89 (4.4%) 43 (2.1%) 49 (2.4%) 117 (5.8%) 1381 (68.2%) 114 (5.6%) 156 (7.7%) 66 (3.3%) 9 (0.4%)	2024 (100%)	0 (0%)
6	settle [factor]	1. Negotiated 2. Imposed 3. None 4. Unclear	304 (15.0%) 123 (6.1%) 1538 (76.0%) 59 (2.9%)	2024 (100%)	0 (0%)
7	maxdur [integer]	Mean (sd) : 149.5 (338.8) min < med < max: 1 < 33 < 4904 IQR (CV) : 165.2 (2.3)	472 distinct values	2024 (100%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
8	mindur	Mean (sd) : 143.9 (338.7)	458 distinct values	2024	0
	[integer]	min < med < max:		(100%)	(0%)
		1 < 26 < 4904			
		IQR (CV) : 154 (2.4)			
9	hiact	1. 1	56 (2.8%)	2024	0
	[ordered, factor]	2. 2	3 (0.1%)	(100%)	(0%)
		3. 3	5 (0.2%)		
		4. 4	4 (0.2%)		
		5. 7	314 (15.5%)		
		6. 8	66 (3.3%)		
		7. 10	19 (0.9%)		
		8. 11	86 (4.2%)		
		9. 12	117 (5.8%)		
		10. 13	16 (0.8%)		
		[9 others]	1338 (66.1%)		
10	hostlev	1. 2	68 (3.4%)	2024	0
	[ordered, factor]	2. 3	602 (29.7%)	(100%)	(0%)
		3. 4	1261 (62.3%)		
		4. 5	93 (4.6%)		
11	recip	1. FALSE	1114 (55.0%)	2024	0
	[logical]	2. TRUE	910 (45.0%)	(100%)	(0%)
12	numa	Mean (sd) : 1.2 (1.4)	13 distinct values	2024	0
	[integer]	min < med < max:		(100%)	(0%)
		1 < 1 < 38			
		IQR (CV) : 0 (1.2)			
13	numb	Mean (sd) : 1.2 (1.2)	13 distinct values	2024	0
	[integer]	min < med < max:		(100%)	(0%)
		1 < 1 < 38			
		IQR (CV) : 0 (1)			
14	ongo2010	1. FALSE	2007 (99.2%)	2024	0
	[logical]	2. TRUE	17 (0.8%)	(100%)	(0%)
15	deaths	1. FALSE	1603 (79.2%)	2024	0
	[logical]	2. TRUE	421 (20.8%)	(100%)	(0%)

MIDB Appendix: Complete Summary

Data Frame Summary

MID_Actor

Dimensions: 4459 x 9
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	V1	Mean (sd) : 2836.8 (1594.5)	4459 distinct values	4459	0
	[integer]	min < med < max:		(100%)	(0%)
		1 < 2902 < 5511			
		IQR (CV) : 2688 (0.6)			

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
2	sidea [integer]	Mean (sd) : 0.5 (0.5) min < med < max: 0 < 1 < 2 IQR (CV) : 1 (1)	0 : 2217 (49.7%) 1 : 2240 (50.2%) 2 : 2 (0.0%)	4459 (100%)	0 (0%)
3	revstate [integer]	Min : 0 Mean : 0.4 Max : 1	0 : 2641 (59.2%) 1 : 1818 (40.8%)	4459 (100%)	0 (0%)
4	revtype1 [integer]	Mean (sd) : 0.7 (1) min < med < max: 0 < 0 < 4 IQR (CV) : 2 (1.3)	0 : 2642 (59.2%) 1 : 607 (13.6%) 2 : 1058 (23.7%) 3 : 124 (2.8%) 4 : 28 (0.6%)	4459 (100%)	0 (0%)
5	fatality [integer]	Min : 0 Mean : 0.2 Max : 1	0 : 3588 (80.5%) 1 : 871 (19.5%)	4459 (100%)	0 (0%)
6	hiact [integer]	Mean (sd) : 9.6 (7.3) min < med < max: 0 < 11 < 22 IQR (CV) : 16 (0.8)	20 distinct values	4459 (100%)	0 (0%)
7	hostlev [integer]	Mean (sd) : 3 (1.3) min < med < max: 1 < 3 < 5 IQR (CV) : 3 (0.4)	1 : 1174 (26.3%) 2 : 191 (4.3%) 3 : 1048 (23.5%) 4 : 1754 (39.3%) 5 : 292 (6.6%)	4459 (100%)	0 (0%)
8	orig [integer]	Min : 0 Mean : 0.9 Max : 1	0 : 524 (11.8%) 1 : 3935 (88.2%)	4459 (100%)	0 (0%)
9	last_days [integer]	Mean (sd) : 163.5 (366.8) min < med < max: 1 < 35 < 4903 IQR (CV) : 179 (2.2)	548 distinct values	4459 (100%)	0 (0%)