MWSUG 2018 - Paper HS094

Genocide Modeling - Historical Risk Factors and Odds Ratios

David J Corliss, Peace-Work, Plymouth, MI

ABSTRACT

This analysis identifies risk factors associated with genocide events. A review of historical conflicts where genocide was present in some and not others provided the data. Using these data, Decision Tree and Random Forest models identify variables with measurable association with genocide events. Logistic Regression and Decision Tree methods are applied to the screened list of variables. Odds ratios are calculated to assess the relative risk of different factors. These models are used to assess the relative likelihood of genocide occurring or developing in the near year in various countries.

INTRODUCTION – GENOCIDE RESEARCH

The term “Genocide” was first used in 1943 in reference to the Armenian Genocide by the Ottoman Empire in 1915-1917. While Genocide research has been very active since that time, an important statistical problem has developed: while case studies have been performed, most previous research has focused on instances where genocide occurred but not where it could have occurred but did not. A historical review of genocide literature finds:

- Genocide research is relatively new
- Previous work has largely focused on reports and case studies
- This had led to a statistical problem: a lack of Control records, where genocide did not occur, prevents rigorous analytic study

There is a need to study cases where genocide did not occur and compare them to instances of genocide. This paper seeks to address this issue in order to identify candidate risk factors.

METHODOLOGY

The methodology used in this study is as follows:

- Identify on-going and recent cases of genocide and the country that perpetrated this crime against humanity
- For each perpetrator country, the most similar country is identified - confronting the same challenges and choices but did not choose to go down this path. Examples.
- Identify potential data sources: country-by-country data
- Eliminate unpromising variables using Bootstrapped Decision Tree
- Remaining variables tested using Single Variable Models
- Odds Ratio plays an important role in identifying potential risk factors
- Investigate the risk factors proposed by this process for reasonableness in connection with being associated with genocide.

DATA SOURCES

Genocide cases from historical sources, including Genocide Watch www.genocidewatch.org and Center for System Peace www.systemicpeace.org.

Country data sources contributing to this study include:

- CIA World Factbook www.cia.gov/library

- World Bank https://data.worldbank.org
- Freedom House https://freedomhouse.org
- PISA Education Survey

CANDIDATE VARIABLE SELECTION USING BOOTSTRAPPED DECISION TREE

A Bootstrapped Decision Tree is an ensemble learning technique for selecting variables to model development. The mathematical framework of this technique was developed by Leo Breiman and Adele Cutler in 2001, who trademarked a name for their implementation. The random selection of candidate model factors uses Tin Kam Ho's Random Subspace method (1995) as a means of stochastic discrimination (E. Kleinberg, 1996).

In this paper, a Bootstrapped Decision Tree is employed for variable selection, identifying and eliminate unintelligent variables from a large number of initial candidate variables. Candidates for subsequent modeling are identified by selecting variables consistently appearing at the top of decision trees created using a random sample of all possible modeling variables. This technique can reduce hundreds of potential predictor fields to a "short list" of 30–50 to be used in developing a model. The process is as follows:

- Use PROC CONTENTS to create a variable list
- Select a random subset of variable names
- Run a decision tree with the selected variables
- Capture the name of the variable selected for the first split – this variable gets one "vote"
- Repeat many times (e.g., 10,000), with the variable at the top of the decision tree getting one vote each time
- Rank the candidate variables by the number of votes received by each

```
* Run a sample with just a few iterations as a test;

%bdt(genocide,genocide_ana,12,10);

* Write the log to a file - needed for Bootstrapping;
PROC PRINTTO LOG='C:\PeaceWork\Genocide\bootstrap_log.log' NEW;
RUN;

* Final run using a large number of iterations;
%bdt(genocide,genocide_ana,12,10000);

PROC PRINTTO;
RUN;
```

The complete source code for this macro in found in the paper "Model Variable Selection Using Bootstrap Decision Tree", David J. Corliss, Proceedings SAS Global Forum 2014.

SINGLE VARIABLE ENSEMBLE MODELING

Once clearly uninformative variables are eliminated using Bootstrapped Decision Tree or some other method, the candidate variables are evaluated individually using single-variable models. An ensemble method has been used, modeling each candidate variable using PROC SURVEYLOGISTIC, PROC

LOGISTIC, and PROC SURVEYREG. The statistical output from all three models is evaluated to determine whether a given candidate variable is a likely risk factor for genocide. As all three modelling methods are applied to each variable to be tested, a macro has been written to apply the three types of models:

```
%macro ensemble(var_name);
proc surveylogistic data=pw.genocide_ana;
   model PerpInd(event='1') = &var_name. / link=probit;
   output out=work.sl_probit p=prob;
run;

proc logistic data=pw.genocide_ana plots=all;
   model PerpInd(event='1') = &var_name.;
   oddsratio &var_name.;
run;

proc surveyreg data=pw.genocide_ana;
   model PerpInd = &var_name.;
run;
%mend;

%ensemble(NE_TRD_GNFS_ZS);
```

RESULTS: POTENTIAL GENOCIDE RISK FACTORS

HUMAN RIGHTS VIOLATIONS

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	53.266	46.327
SC	54.877	49.549
-2 Log L	51.266	42.327

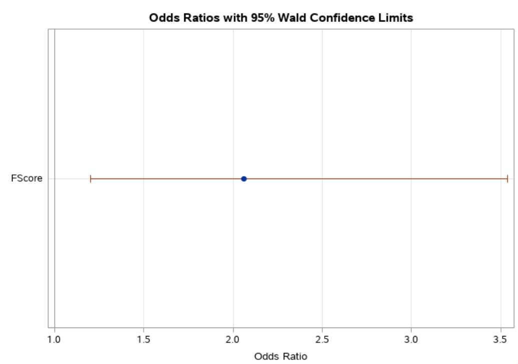
Testing Global Null Hypothesis: BETA=0				
Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	8.94	1	36	0.0050
Score	9.65	1	36	0.0037
Wald	6.43	1	36	0.0157

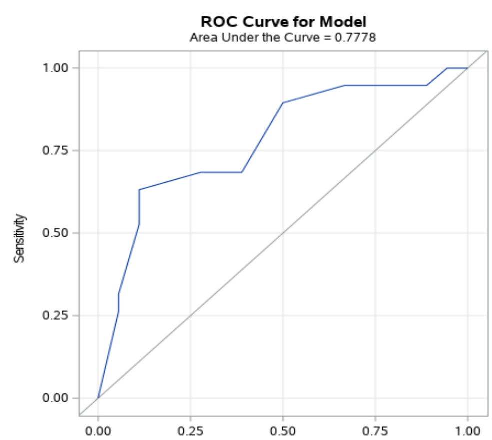
Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-1.9865	0.8514	-2.33	0.0253
FScore	0.4322	0.1704	2.54	0.0157

NOTE: The degrees of freedom for the t tests is 36.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.3372	1.3369	6.2317	0.0125
FScore	1	0.7239	0.2753	6.9163	0.0085

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
FScore	2.062	1.202	3.537





SERVICE SECTOR % OF GDP

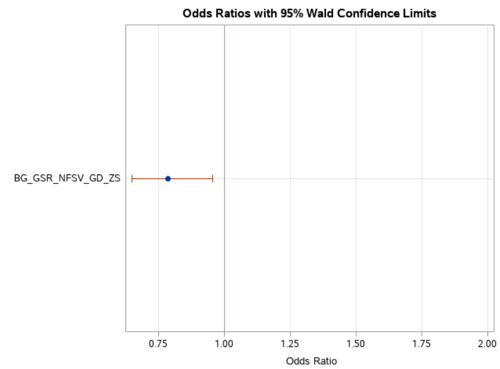
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	40.673	28.414
SC	42.005	31.078
-2 Log L	38.673	24.414

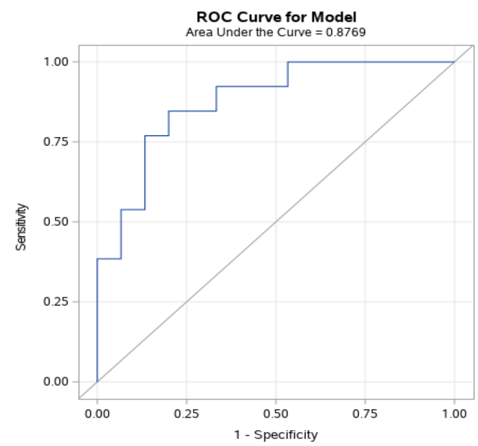
Testing Global Null Hypothesis: BETA=0				
Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	14.26	1	27	0.0008
Score	98.41	1	27	<.0001
Wald	8.25	1	27	0.0078

Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.9220	0.6449	2.98	0.0060
BG_GSR_NFSV_GD_ZS	-0.1413	0.0492	-2.87	0.0078
NOTE: The degrees of freedom for the t tests is 27.				

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.2272	1.3229	5.9515	0.0147
BG_GSR_NFSV_GD_ZS	1	-0.2402	0.0992	5.8693	0.0154

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
BG_GSR_NFSV_GD_ZS	0.786	0.648	0.955





IMPORT / EXPORT % OF GDP

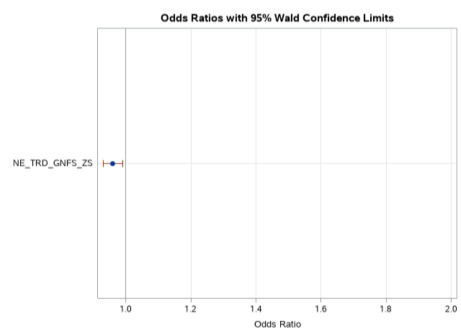
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	50.492	42.400
SC	52.047	45.511
-2 Log L	48.492	38.400

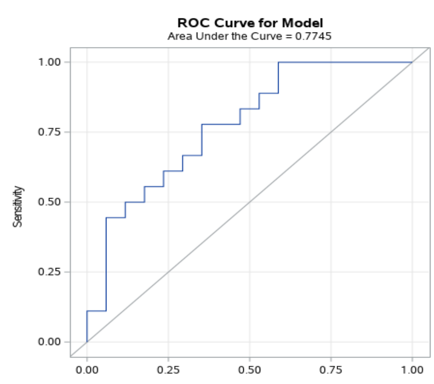
Testing Global Null Hypothesis: BETA=0				
Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	10.09	1	34	0.0032
Score	24.52	1	34	<.0001
Wald	12.53	1	34	0.0012

Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1.7582	0.5710	3.08	0.0041
NE_TRD_GNFS_ZS	-0.0256	0.00722	-3.54	0.0012
NOTE: The degrees of freedom for the t tests is 34.				

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.8524	1.1065	6.6453	0.0099
NE_TRD_GNFS_ZS	1	-0.0414	0.0160	6.6859	0.0097

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
NE_TRD_GNFS_ZS	0.959	0.930	0.990





IMPORT / EXPORT % OF GDP

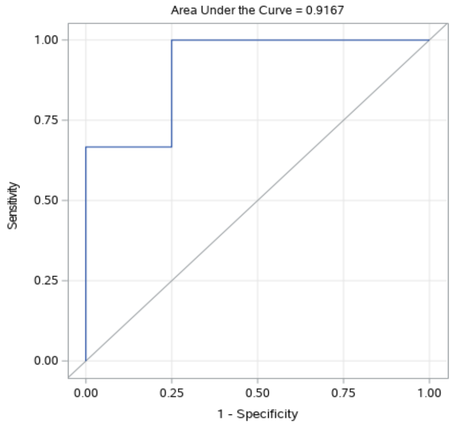
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	15.460	10.548
SC	15.763	11.153
-2 Log L	13.460	6.548

Testing Global Null Hypothesis: BETA=0				
Test	F Value	Num DF	Den DF	Pr > F
Likelihood Ratio	6.91	1	9	0.0274
Score	121.43	1	9	<.0001
Wald	8.80	1	9	0.0158

Analysis of Maximum Likelihood Estimates				
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	2.3213	1.1165	2.08	0.0674
CM_MKT_LCAP_GD_ZS	-0.0387	0.0130	-2.97	0.0158
NOTE: The degrees of freedom for the t tests is 9.				

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	3.9466	2.2367	3.1134	0.0777
CM_MKT_LCAP_GD_ZS	1	-0.0646	0.0403	2.5691	0.1090

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
CM_MKT_LCAP_GD_ZS	0.937	0.866	1.014



CONCLUSION

This investigation proposes a rigorous and effective methodology of the statistical analysis of genocide events, identifying candidate risk factors in a reproducible manner. This is enabled by paring perpetrator countries with highly similar countries facing as much of the same context and challenges as possible, but not implicated in genocide. Country-by-Country data from government agencies and NGOs provides candidate predictors.

Genocide risk factors range from traditional contributors such as human rights violations to more subtle socio-economic indicators including weak services and import/export sectors, low market capitalization of publicly traded companies, and an absence of PISA data. Applying these risk factors to the population of all countries (excluding microstates), this methodology indicates countries at risk of genocide events in the near- to mid-term future include Eritrea, Western Sahara, and Guinea.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

David J Corliss, PhD
Peace-Work
davidjcorliss@peace-work.org
www.peace-work.org