Reproducible Statistical Analysis Supporting Thesis

John G Schwitz

February 27, 2014

1 Introduction

This section assures that all the analysis is reproducible. The data, data processing, and statistical approaches are documented. The data is from: Beyond Greed and Grievance: Feasibility and Civil War by Paul Collier, Anke Hoeffler, and Dominic Rohner. Oxford Economic Papers (2008). The dataset is attainable from: http://www.apsanet.org/content_29436.cfm.

The dependent variable is a binary indicator on the onset of war: warglediteh_july2006.

1.1 Statistical Approach

For ordinary regression the variables are fitted by minimizing the sum of the squared distance of individual points (observations) to a line. This approach does not work for variables related non-linearly as they are in logistics regression. In this case the variables are first transformed to form a linear relationship and a different method of fit, maximum likelihood, is used.

Logistic regression was used in all the models of fragile states discussed in this paper. This non-linear approach is used when the dependent variable is binary (yes/no). For fragile states the challenge is to forecast the probability of the onset of an political instability event. Instability versus stability is a binary event. As in betting, the odds ratio is the probability of the event divided by the probability of the event not occurring. The dependent variable in logistic regression is the logarithm of the odds ratio of instability occurring.

The logarithm of the odds ratio of instability transforms the equation into linear form enabling regression on independent variables. Regression is then performed using maximum likelihood with a solution that encompasses the concept of confidence intervals. However; point forecasts are insufficient for analysis. Statistical approaches also provide ranges and confidence levels for a forecast. For instance, 95 percent confidence level implies that with 95 percent certainty the dependent variable (onset of violence) is within the designated range.

Using maximum likelihood to fit a non-linear regression invalidates R Squared as a measure of fit (it cannot be calculated). A plethora of pseudo R Square measures, all suspect, have been employed to remedy this. Best statistical practice for non-linear regression is using a measure that trades off fit against the number of model parameters. One such measure is the Akaike information criterion (AIC) which rewards goodness of fit while also penalizing the use of estimation parameters. This protects against overfitting because it is always possible to improve fit through adding parameters (overfitting) without contributing to the explanatory power of the model. The logistic regressions in the thesis employ this approach.

2 Collier and Hoeffler 2009 Analysis

2.1 Load Collier 2009 dataset

The analysis uses the original Collier 2009 dataset which consists of 1872 observations of 40 variables. Each observation encompasses a 5 year period and span from 1965 to 2005. Observations with continuing conflict are labeled NA. Afghanistan in 1980, 1985, 1990, 1995, 2000 is classifed as NA. Data processing is performed to eliminate any binary indicators for war (labeled NA). This results in 1063 observations of 40 variables reproducing the Collier base case.

```
##  Data
data <- read.dta("../Data/collier and hoeffler 2009.dta")
##  str(data)  ##  1872 obs. of 40 variables

##  From Yonamine in analysis of 2004 data.
##  data<-data[complete.cases(data$wargleditsch_july2006),]
##  this drops all rows with NA for wars,
##  It drops all continuations of wars C&H use this approach
##  str(data)  ##  1658 obs. of 40 variables
##  examine data that has war binary NA
##  data.out<-data[(is.na(data$wargleditsch_july2006)==TRUE),]
##  Afghanistan 1980,85,90,95,00 is in this
##  However; Collier Base Case indicator is sample8july

##  Eliminate dependent variables not used in analysis
##  This enables Amelia algorithm to work NA's
```

2.2 Trim dataset to relevant elements for Collier Base Case

```
data <- remove.vars(data, c("years_since_indep","postcoldwar","pop","priostart3end1",
      "distusa","previouswar_wargleditschex","colfra2","ethfrac_mean","td3","td4",
      "td5","td6","td7","td8","td9","firstwar8july06","xerfracdom",
      "popd_amean","fuel_amean","distcr","tropicar","sample911"))
```

```
Removing variable 'years_since_indep'
Removing variable 'postcoldwar'
Removing variable 'pop'
Removing variable 'priostart3end1'
Removing variable 'distusa'
Removing variable 'previouswar_wargleditschex'
Removing variable 'colfra2'
Removing variable 'ethfrac_mean'
Removing variable 'td3'
Removing variable 'td4'
Removing variable 'td5'
Removing variable 'td6'
Removing variable 'td7'
Removing variable 'td8'
Removing variable 'td9'
Removing variable 'firstwar8july06'
Removing variable 'xerfracdom'
Removing variable 'popd_amean'
Removing variable 'fuel_amean'
Removing variable 'distcr'
Removing variable 'tropicar'
Removing variable 'sample911'
```

```
##  Dependent variables reduced from 40 to 18
##  1872 observations

##  Tailor dataset to Collier Base Case (Selected on sample8july)
##  Eliminate rows with wargleditsch_july2006 NA
##  this drops all rows with NA for wars, i.e. drops all continuations of conflict.
##  Collier used this approach, which is reasonable
data.waronset <-data[complete.cases(data$wargleditsch_july2006),]
##  str(data)  ##  1658 obs. of 18 variables

##  Select on Collier data sample (Indicator sample8july)
data.collier <- data[complete.cases(data$sample8july),]
##  str(data.collier)  1063 obs. of 18 variables -- As in paper
```

2.3 Predictor Variables in Collier Base Case

Predictor Variable	Abbreviated Name	Variable
Log GDP per capita	LogGDPperCap	lnnewgdp
Growth in GDP per capita	ChngGDP	newgrowth
GDP from commodity exports	ComXport	sxp_may06
GDP from commodity exports SQ	ComXportSQ	sxp2_may06
Years of Peace	Years Peace	peace_gleditsch
Former French Colony in Africa	FrAfCol	colfra_af
Social Fractionalization	Frac	xerfrac_march08_mean
Percent young men in population	YMen	ymen1529_march08_mean
Log of population	LogPop	lnpop
Percent mountainous terrain	Mount Terrain	mount_march08_mean

Table 1: Collier Predictor Variables

2.4 Logit Analysis from Collier Base Case

```
##  logit analysis from Collier Base Case
z.out.collier <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth + sxp_may06 +
                sxp2_may06 + peace_gleditsch + colfra_af + xerfrac_march08_mean +
                ymen1529_march08_mean + lnpop + mount_march08_mean,
                model = "logit", data=data.collier )
```

```
Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

```
 How to cite this model in Zelig:
  Kosuke Imai, Gary King, and Olivia Lau. 2014.
  "logit: Logistic Regression for Dichotomous Dependent Variables"
  in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software,"
  http://gking.harvard.edu/zelig
```

```
##  Matches Collier Base Case (Table 3 Column 4 p. 9)

##  summary(z.out.collier)
##  Only 3 digits
options(digits = 3)
##  Collier means
x.collier.mean <- setx(z.out.collier)

##  Simulate for Collier and Rare Imputed with dependent variables at means
s.collier.mean <- sim(z.out.collier, x.collier.mean)

##  Collier means
comparematrix <- cbind(c("CollierAvg", x.collier.mean$values))
```

2.5 Successfully duplicated Collier Base Case

At this point the thesis has successfully replicated the Collier base case. The stars on the p value are a measure quantifying the acceptance of the value of the dependent variable and rejecting the null hypothesis (the value is zero). For instanse Years Peace having no effect (zero) has only .0001 percent probability).

The next page presents the range of uncertainty around this estimate.

	Collier Base Model
LogGDPperCap	-0.22 (0.12)
chngGDP	-0.14 (0.04)***
ComXport	6.99 (3.95)
ComXportSQ	-14.44 (7.93)
Years Peace	-0.06 (0.01)***
FrAfCol	-1.22 (0.61)*
Frac	2.19 (0.81)**
YMen	12.64 (8.17)
LogPop	0.27 (0.10)**
Mount Terrain	0.01 (0.01)
AIC	399.61
Log Likelihood	-188.80
Deviance	377.61
Num. obs.	1063

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

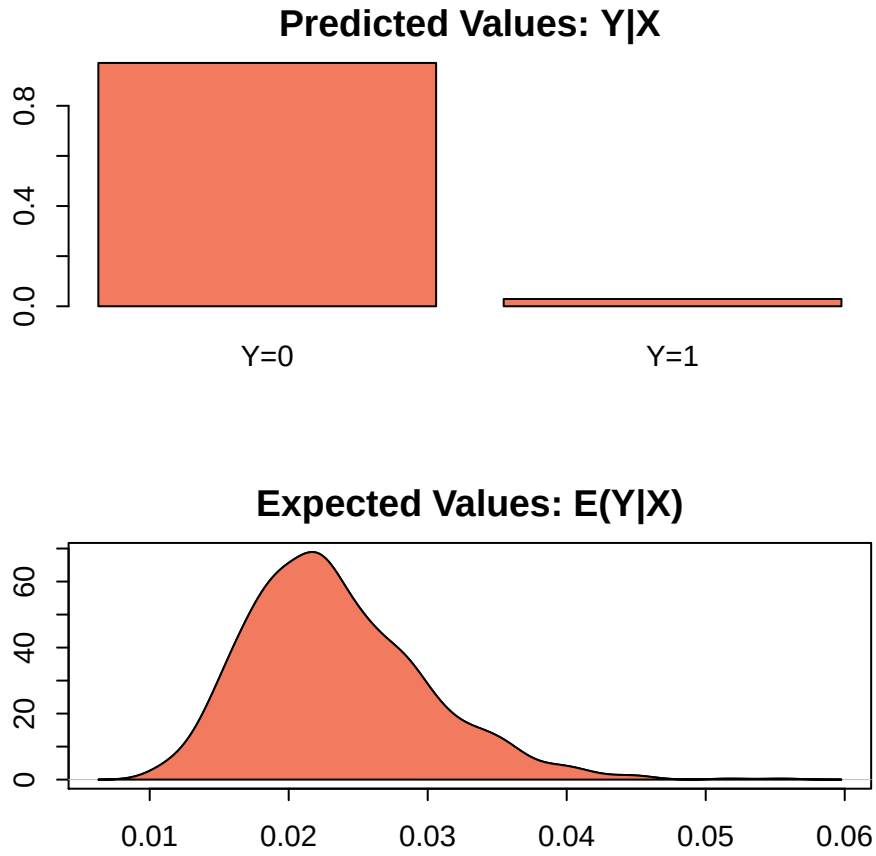
Table 2: Duplicate Collier Results

2.6 Uncertainty of Collier estimate

An additional step is taken to portray the uncertainty of conflict in a manner understandable to the policymaker. Both the predicted value of conflict and the range of uncertainty around this point estimate is plotted below. The average result is a point estimate for conflict; however, this is presented with lower and upper bounds which encompass a 95 percent probability (2.5 percent at each tail).

Table 3: Uncertainty in Collier Base Case

```
s.collier.mean$stats[1]
$`Expected Values: E(Y|X)`
   mean     sd    50%   2.5%  97.5%
 0.0234 0.0063 0.0225 0.0137 0.0376
plot(s.collier.mean)
```



2.7 Logit Analysis with imputed Missing Values

The Amelia imputation program is used to provide estimates of missing values.

```
##  Examine rows not used in Collier Analysis
##  data.elim<-data[(is.na(data$wargleditsch_july2006)==TRUE),]
##  All these have NA's
##  List to discover
##  data.elim[1:10,c(2,3,14)]

##  Imputations on NA
##  This increases dataset from Collier 1063 to 1658 observations
data.impute <- amelia(data.waronset, m = 1, ts = "year", cs = "country")
```

```
-- Imputation 1 --

  1  2  3  4  5  6  7  8  9
```

```
data.impute <-data.impute$imputations[[1]]
##  name(data.impute)  ##  Provides column names

##  Demonstrate sensitivity of parameters to data
##  Run regression with imputed larger dataset
##  logit analysis from Collier Base Case
z.out.impute <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth +
  sxp_may06 + sxp2_may06 + peace_gleditsch + colfra_af +
        xerfrac_march08_mean + ymen1529_march08_mean + lnpop +
        mount_march08_mean,
        model = "logit", data=data.impute )
```

```
 How to cite this model in Zelig:
  Kosuke Imai, Gary King, and Olivia Lau. 2014.
  "logit: Logistic Regression for Dichotomous Dependent Variables"
  in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software,"
  http://gking.harvard.edu/zelig
```

2.8 Results for Logit Analysis with Imputed Values

The R statistical package has powerful algorithms to impute data values for missing values (Amelia). However; the dataset must be examined because if data is missing for all years for a country the imputed values are not valid.

This is the case for GDP for 20 countries with others missing almost all values.

This problem is also encountered with commodity exports as a percentage of GDP. This problem then also impacts the squared value which is another dependent variable.

In addition to these severe problems the AIC is far worse that the Collier Base Case. Therefore; the imputed logit analysis is not used.

```
##  Missing data causes imputations to vary significantly from Collier
##  Imputation means
x.impute.mean <- setx(z.out.impute)
##  Compare Collier results with results from imputed dataset with means
##  Better variable names
c0 <- c("Case", "(Intercept)", "LogGDPperCap", "chngGDP",
                "ComXport", "ComXportSQ", "Years Peace", "FrAfCol",
                "Frac", "YMen", "LogPop", "Mount Terrain")
c1 <- c("CollierAvg",NA,x.collier.mean$values)
c2 <- c("ImputeAvg",NA,x.impute.mean$values)
df <- data.frame(cbind(c0, c1, c2))
```

```
Warning:  some row.names duplicated:  2 -> row.names NOT used
```

	Model with imputed values
LogGDPperCap	-0.26 (0.10)**
chngGDP	-0.09 (0.03)**
ComXport	-0.75 (1.55)
ComXportSQ	0.53 (1.59)
Years Peace	-0.05 (0.01)***
FrAfCol	-1.14 (0.57)*
Frac	2.23 (0.71)**
YMen	3.40 (6.63)
LogPop	0.29 (0.07)***
Mount Terrain	0.01 (0.01)
AIC	563.27
Log Likelihood	-270.63
Deviance	541.27
Num. obs.	1658

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 4: Model with imputed values

2.9 Comparison of Collier versus Impute Means

A comparison of Collier versus imputed means. The Collier means are interpreted as the average value for each of the dependent variables from the Collier Base Case sample of 1063 observations.

All comparisons use these values until explicitly changed.

Table 5: Comparison of Collier versus Base Case Means

```
##  Compare Collier results with results from imputed dataset with means
df
               c0         c1        c2
1            Case CollierAvg ImputeAvg
2     (Intercept)         NA        NA
3    LogGDPperCap       7.59      7.59
4         chngGDP       1.84      2.18
5        ComXport      0.164     0.182
6      ComXportSQ     0.0598    0.0758
7     Years Peace         32      30.1
8         FrAfCol      0.101     0.073
9            Frac       0.18     0.175
10           YMen      0.129     0.129
11         LogPop       15.4      14.9
12 Mount Terrain       15.8      16.8
```

2.10 Results Comparison of Collier versus Impute Models

A side by side comparison of the Collier Base Case and the results of the imputation. There are several significant differences in GDP growth, commodity exports, and percentage of young men.

The AIC is significantly higher in the imputed model which is a signal that the Collier Base Case is a better model.

	Collier Base Model	Impute Model
LogGDPperCap	−0.22 (0.12)	−0.26 (0.10)**
chngGDP	−0.14 (0.04)***	−0.09 (0.03)**
ComXport	6.99 (3.95)	−0.75 (1.55)
ComXportSQ	−14.44 (7.93)	0.53 (1.59)
Years Peace	−0.06 (0.01)***	−0.05 (0.01)***
FrAfCol	−1.22 (0.61)*	−1.14 (0.57)*
Frac	2.19 (0.81)**	2.23 (0.71)**
YMen	12.64 (8.17)	3.40 (6.63)
LogPop	0.27 (0.10)**	0.29 (0.07)***
Mount Terrain	0.01 (0.01)	0.01 (0.01)
AIC	399.61	563.27
Log Likelihood	-188.80	-270.63
Deviance	377.61	541.27
Num. obs.	1063	1658

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6: Model comparison Collier and Impute

2.11 Rare Logit Analysis on Collier dataset

A special approach to rare data is applied which also creates an unbiased estimate. This approach is applied to the Collier Base observations of 1063.

Since the AIC's are identical the rare data result will be used for all future variations.

```
##  Missing data causes imputations to vary significantly from Collier Analysis
##  Rare Data Run without imputation
##  Need to calculate Tau percent of sample with conflict
##  Collier has 1063 observations with 38 state of war .038 is Tau
##
z.out.rare <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth +
  sxp_may06 + sxp2_may06 + peace_gleditsch + colfra_af +
        xerfrac_march08_mean + ymen1529_march08_mean + lnpop +
        mount_march08_mean, model = "relogit", tau = .038, data=data.collier )
```

```
Warning:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

```
 How to cite this model in Zelig:
  Kosuke Imai, Gary King, and Olivia Lau. 2014.
  "relogit: Rare Events Logistic Regression for Dichotomous Dependent Variables"
  in Kosuke Imai, Gary King, and Olivia Lau, "Zelig: Everyone's Statistical Software,"
  http://gking.harvard.edu/zelig
```

```
##  Imputation means
x.rare.mean <- setx(z.out.rare)
```

2.12 Results Comparison of Collier versus Rare

Collier Base versus results for Rare logistics regression.

	Collier	Rare
LogGDPperCap	−0.22 (0.12)	−0.21 (0.12)
chngGDP	−0.14 (0.04)***	−0.14 (0.04)***
ComXport	6.99 (3.95)	5.90 (3.91)
ComXportSQ	−14.44 (7.93)	−11.67 (7.85)
Years Peace	−0.06 (0.01)***	−0.05 (0.01)***
FrAfCol	−1.22 (0.61)*	−1.09 (0.60)
Frac	2.19 (0.81)**	2.14 (0.80)**
YMen	12.64 (8.17)	13.39 (8.08)
LogPop	0.27 (0.10)**	0.25 (0.10)**
Mount Terrain	0.01 (0.01)	0.01 (0.01)
AIC	399.61	399.61
Log Likelihood	-188.80	-188.80
Deviance	377.61	377.61
Num. obs.	1063	1063

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7: Model comparison Collier and Rare

2.13 Simulation of Collier and Rare Data at Mean

This is completes the first of three stages in the analysis. The second stage is to apply this forecast to a state and run scenarios. The third stage is to apply the percolation model.

This portrays the results of first the Collier base regression then the rare regression on the means of the dependent variables for the population. The predicted value of 1 is for the onset of violence. This provides a base for comparison.

Table 8: Collier Base Case at the Average

```
##  Collier Simulation
s.collier.mean$stats
$`Expected Values: E(Y|X)`
   mean     sd    50%   2.5%  97.5%
 0.0234 0.0063 0.0225 0.0137 0.0376

$`Predicted Values: Y|X`
     0     1
 0.971 0.029

attr(,"class")
[1] "summarized.qi"
```

Table 9: Rare case at the Average

```
##  Rare Simulation
s.rare.mean <- sim(z.out.rare, x.rare.mean)
s.rare.mean$stats
$`Expected Values: E(Y|X)`
   mean      sd    50%    2.5% 97.5%
 0.0156 0.00433 0.0152 0.00872 0.026

$`Predicted Values: Y|X`
     0     1
 0.981 0.019

attr(,"class")
[1] "summarized.qi"
```

2.14 Afghanistan 2012 Collier Logit

First comparison is Afghanistan 2012 Collier versus Rare models.

Investigate model prediction for Afghanistan. The three scenarios are Afghanistan with 2012 data, and two cases from the World Bank publication; Afghanistan in Transition: Looking Beyond 2014 Volume 2: Main Report (May 2012).

The World Bank predictions encompass average values for 2011 - 2019. The scenarios are the World Bank Base Case and a Deteriorating Security Case.

Point estimates and uncertainty are portrayed on the next pages.

```
##  Afghanistan with 2012 data Collier Model
x.afghan.2012C <- setx(z.out.collier, ymen1529_march08_mean=.1288071,
              xerfrac_march08_mean=.1789208, mount_march08_mean=65.60001,
              wargleditsch_july2006=1, africa=0, lnnewgdp=6.0331, newgrowth=.117,
              lnpop=17.211, peace_gleditsch=0, sxp_may06=.33, sxp2_may06=.1089,
              colfra_af=0)
s.afghan.2012C <- sim(z.out.collier, x.afghan.2012C)

##  Afghanistan with 2012 data Rare Model
x.afghan.2012R <- setx(z.out.rare, ymen1529_march08_mean=.1288071,
              xerfrac_march08_mean=.1789208, mount_march08_mean=65.60001,
              wargleditsch_july2006=1, africa=0, lnnewgdp=6.0331, newgrowth=.117,
              lnpop=17.211, peace_gleditsch=0, sxp_may06=.33, sxp2_may06=.1089,
              colfra_af=0)
s.afghan.2012R <- sim(z.out.rare, x.afghan.2012R)

##  Afghanistan at World Bank Base Case
x.afghan.WBbase <- setx(z.out.rare, ymen1529_march08_mean=.1288071,
              xerfrac_march08_mean=.1789208, mount_march08_mean=65.60001,
              wargleditsch_july2006=1, africa=0, lnnewgdp=6.0331, newgrowth=.059,
              lnpop=17.211, peace_gleditsch=0, sxp_may06=.33, sxp2_may06=.1089,
              colfra_af=0)
s.afghan.WBbase <- sim(z.out.rare, x.afghan.WBbase)

##  Afghanistan at World Bank Deteriorating Security
x.afghan.WBDet <- setx(z.out.rare, ymen1529_march08_mean=.1288071,
              xerfrac_march08_mean=.1789208, mount_march08_mean=65.60001,
              wargleditsch_july2006=1, africa=0, lnnewgdp=6.0331, newgrowth=.007,
              lnpop=17.211, peace_gleditsch=0, sxp_may06=.4, sxp2_may06=.1089,
              colfra_af=0)
s.afghan.WBDet <- sim(z.out.rare, x.afghan.WBDet)
```

2.15 Simulation of Collier and Rare Data at 2012

This is examines Collier versus Rare using Afghanistan 2012 data.

This portrays the results of first the Collier base regression then the rare regression on the means of the dependent variables for the population. The predicted value of 1 is for the onset of violence. This provides a base for comparison.

Table 10: Collier Base Case at the Afghanistan 2012

```
##  Collier Simulation
s.afghan.2012C$stats
$`Expected Values: E(Y|X)`
  mean    sd   50%  2.5% 97.5%
 0.545 0.121 0.548 0.304  0.77

$`Predicted Values: Y|X`
     0     1
 0.452 0.548

attr(,"class")
[1] "summarized.qi"
```

Table 11: Rare case at the Afghanistan 2012

```
##  Rare Simulation
s.afghan.2012R$stats
$`Expected Values: E(Y|X)`
  mean    sd   50%  2.5% 97.5%
 0.407 0.118 0.403 0.187 0.644

$`Predicted Values: Y|X`
     0     1
 0.614 0.386

attr(,"class")
[1] "summarized.qi"
```

2.16 Comparison of Afghanistan Scenarios

A comparison of the three scenarios generating in the following pages. The confidence intervals are on 1 standard deviation (sd) which encompasses 68

Scenario	Expectation of Violence	sd	lower bound	upper bound
Afghanistan 2012	41%	12%	29%	52%
World Bank Base	41%	12%	29%	52%
WB Deteriorating Security	50%	16%	34%	66%

Table 12: Afghanistan Cases

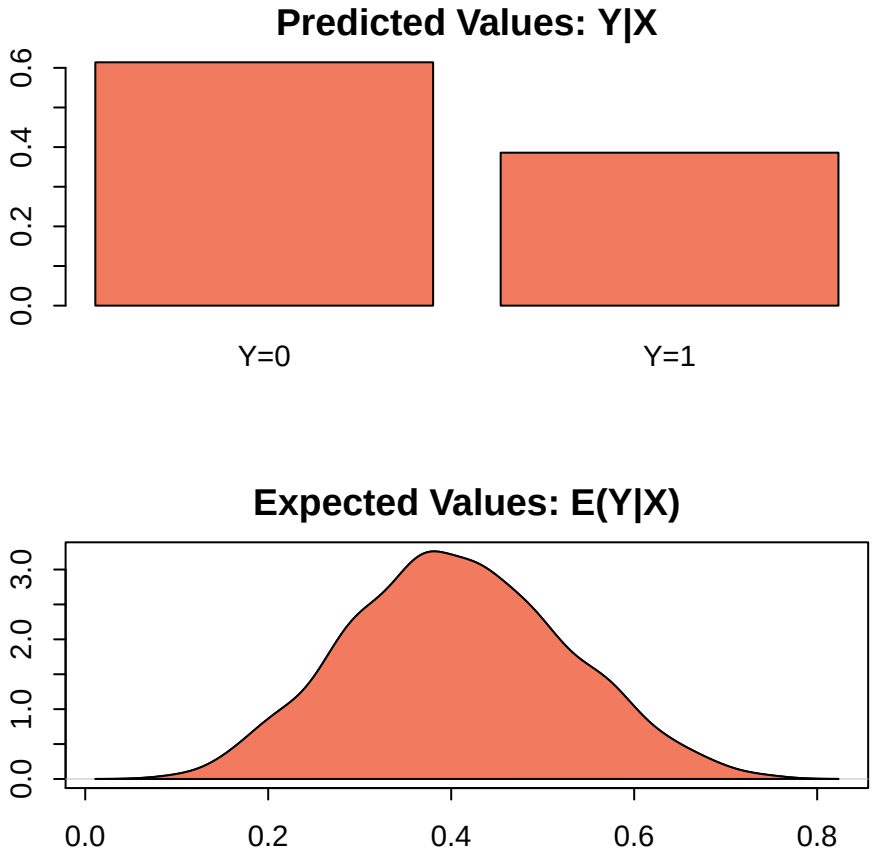
2.17 Afghanistan 2012

This is the predicted probability of the onset of violence using Afghanistan 2012 values for predictor variables.

An additional step is taken to portray the uncertainty of conflict in a manner understandable to the policymaker. Both the predicted value of conflict and the range of uncertainty around this point estimate is plotted below. The range of uncertainty with lower and upper bounds encompass 95 percent certainty (2.5 percent in both the lower and upper tails.

Table 13: Forecast for Afghanistan

```
s.afghan.2012R$stats[1]
$`Expected Values: E(Y|X)`
  mean    sd   50%  2.5% 97.5%
 0.407 0.118 0.403 0.187 0.644
plot(s.afghan.2012R)
```

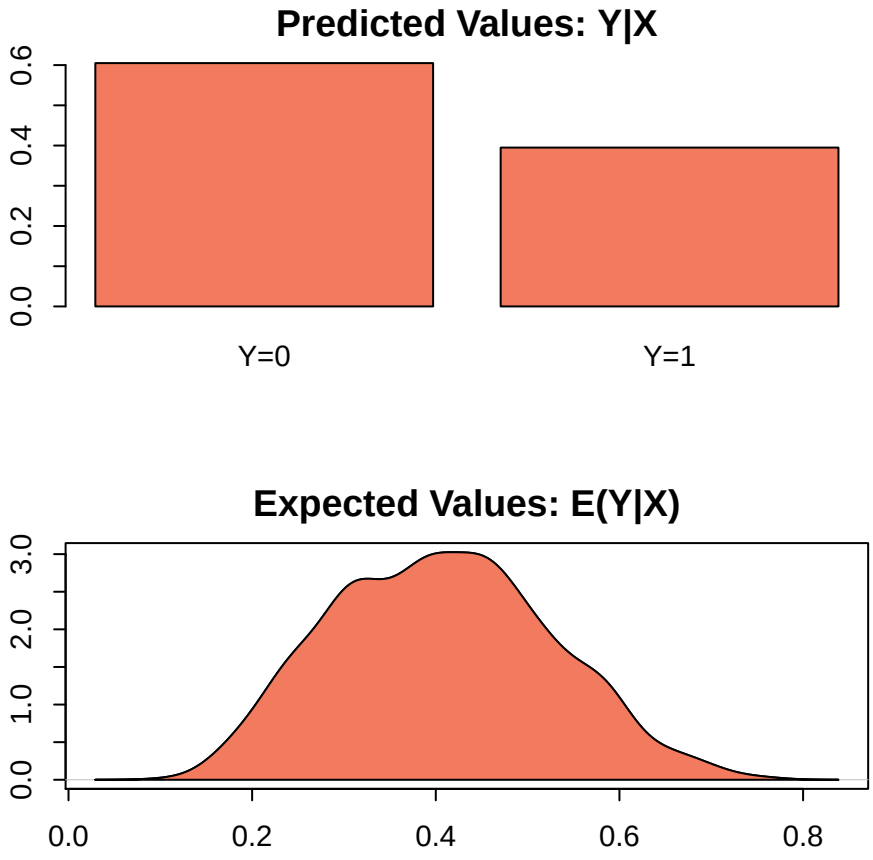


2.18 World Bank Base Case 2011-2019

This is the predicted probability of the onset of violence using the World Bank Base Case for average predictor values from 2011 - 2019.

Table 14: World Bank Base Case

```
s.afghan.WBbase$stats[1]
$`Expected Values: E(Y|X)`
  mean    sd   50%  2.5% 97.5%
 0.407 0.118 0.404 0.196 0.646
plot(s.afghan.WBbase)
```

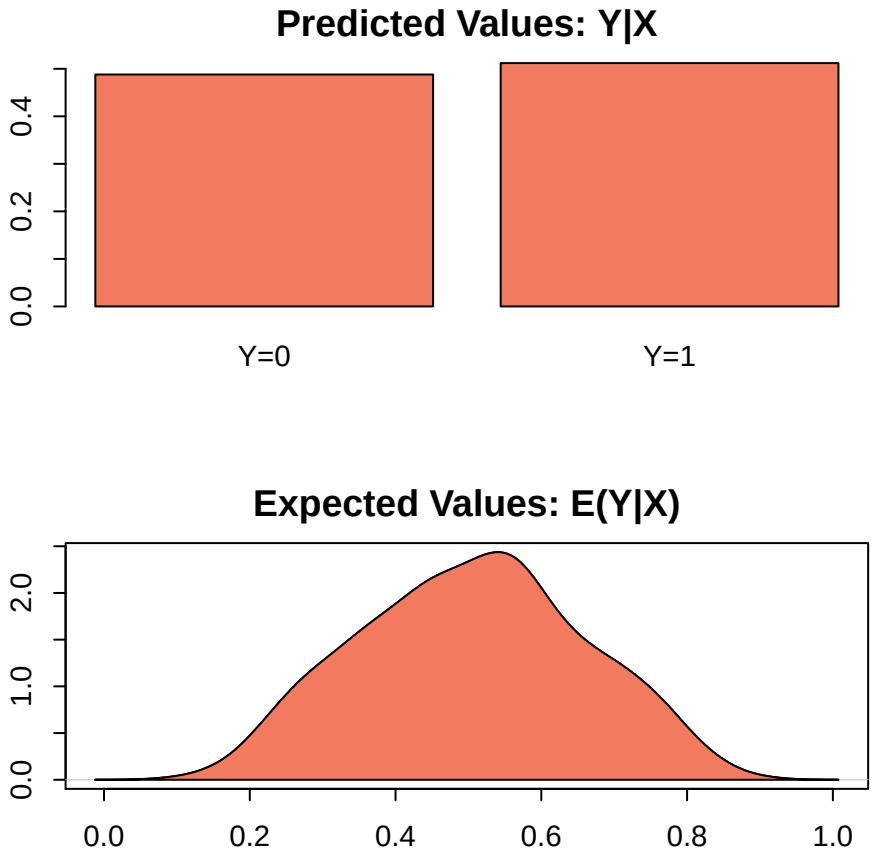


2.19 World Bank Base Deteriorating Security 2011-2019

This is the predicted probability of the onset of violence using the World Bank Deteriorating Security for average predictor values from 2011 - 2019.

Table 15: World Bank Deteriorating Security Case

```
s.afghan.WBDet$stats[1]
$`Expected Values: E(Y|X)`
  mean    sd   50%  2.5% 97.5%
 0.505 0.155 0.509 0.219 0.791
plot(s.afghan.WBDet)
```



2.20 Investigate missing values in imputation

```
##  Investigate missing values
##  Select the set of rows with missing values
data.NA <- data.waronset[!complete.cases(data.waronset),]
##  Examine for missing values

##  For ymen1529_march08_mean (0 Found)
columns <- c(2,3,4)
data.inv <- data.NA[,columns]
data.inv <- subset(data.inv, is.na(data.inv$ymen1529_march08_mean))

##  For xerfrac_march08_mean (0 Found)
columns <- c(2,3,5)
data.inv <- data.NA[,columns]
data.inv <- subset(data.inv, is.na(data.inv$xerfrac_march08_mean))

##  For mount_march08_mean (0 Found)
columns <- c(2,3,6)
data.inv <- data.NA[,columns]
data.inv <- subset(data.inv, is.na(data.inv$mount_march08_mean))

##  For lnnewgdp (351 Found)
columns <- c(2,3,10)
data.inv <- data.NA[,columns]
data.inv <- subset(data.inv, is.na(data.inv$lnnewgdp))
##  data.inv

##  For sxp_may06 (412 Found)
columns <- c(2,3,14)
data.inv <- data.NA[,columns]
data.inv <- subset(data.inv, is.na(data.inv$sxp_may06))
##  data.inv
```

2.21 Investigate Alternative Logistic Regression

```
##  logit analysis from Collier Base Case PLUS Democracy (AIC 401.6)
##  z.out.BPLUS <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth + sxp_may06 +
##                sxp2_may06 + peace_gleditsch + colfra_af + xerfrac_march08_mean +
##                ymen1529_march08_mean + lnpop + mount_march08_mean +
##                democ_amean, model = "logit", data=data.collier )
##  summary(z.out.BPLUS)

##  logit analysis from Collier Base Case PLUS Democracy -M (AIC 401.7)
##  z.out.BM <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth + sxp_may06 +
##                sxp2_may06 + peace_gleditsch + colfra_af + xerfrac_march08_mean +
##                ymen1529_march08_mean + lnpop +
##                democ_amean, model = "logit", data=data.collier )
##  summary(z.out.BM)

##  logit analysis from Collier Base Case PLUS Democracy MINUS (AIC 401.4)
##  z.out.MINUS <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth +
##          peace_gleditsch + colfra_af + xerfrac_march08_mean + lnpop +
##          democ_amean, model = "logit", data=data.collier )
##  summary(z.out.MINUS)

##  logit analysis for NO Mountain (AIC 399.7)
##  z.out.NOM <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth + sxp_may06 +
##          sxp2_may06 + peace_gleditsch + colfra_af + xerfrac_march08_mean +
##          ymen1529_march08_mean + lnpop, model = "logit", data=data.collier )
##  summary(z.out.NOM)

##  logit analysis for NO commodity exports (AIC 401)
##  z.out.NOX <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth +
##              sxp2_may06 + peace_gleditsch + colfra_af + xerfrac_march08_mean +
##              ymen1529_march08_mean + lnpop + mount_march08_mean,
##              model = "logit", data=data.collier )
##  summary(z.out.NOX)

##  logit analysis for NO young men (AIC 399.8)
##  z.out.NOYM <-zelig(wargleditsch_july2006 ~  lnnewgdp + newgrowth + sxp_may06 +
##              sxp2_may06 + peace_gleditsch + colfra_af + xerfrac_march08_mean +
##              lnpop + mount_march08_mean,
##              model = "logit", data=data.collier )
##  summary(z.out.NOYM)
```

2.22 NetLogo Model

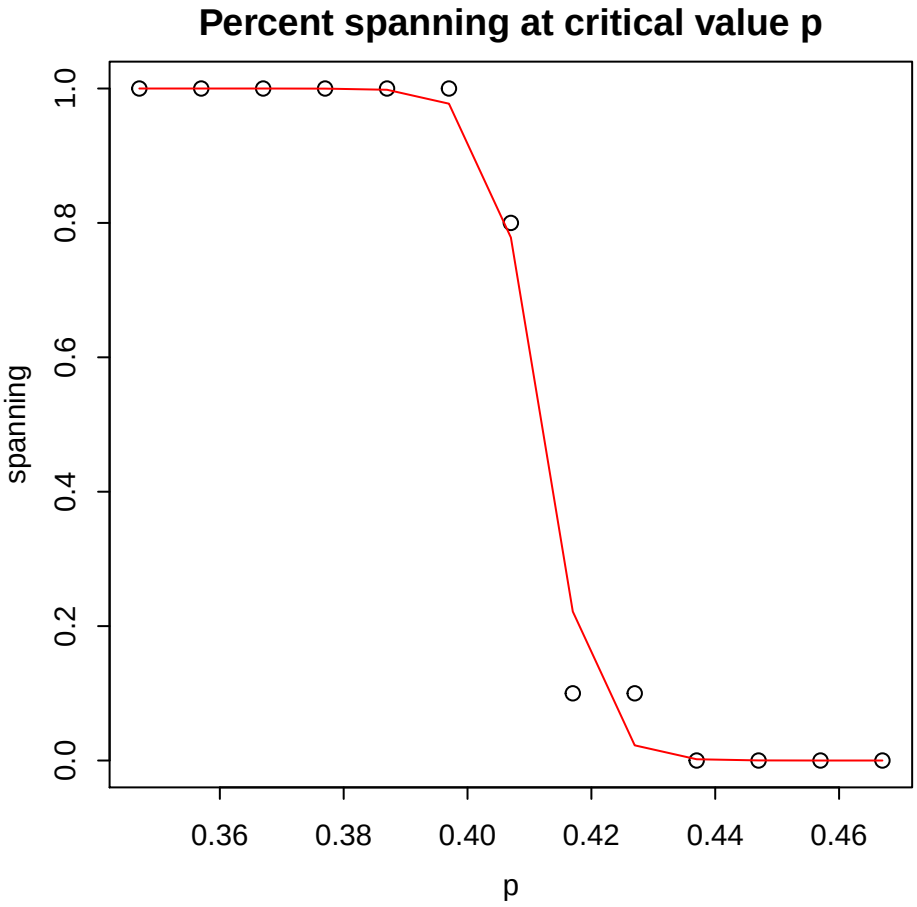
This section documents the NetLogo Model used in the thesis. The model is from: Pattern & Process: Spatial Simulation Exploring Pattern & Process by David O'Sullivan and George Perry. The model used 5.1-sitePercolation-using-R.nlogo is available from: http://patternandprocess.org/related-software-links/

The model was run for iterations of 10 from p=.347 to .467

```
##  From study on Thesis model
##  Plot s curve
spanning <- c(1,1,1,1,1,1,.8,.1,.1, 0, 0, 0, 0)
p <- (c(.347, .357, .367, .377, .387, .397, .407, .417, .427, .437, .447, .457, .467))
crisis <- data.frame(spanning, p)
logistic.out = glm(spanning ~ p, family=binomial(logit), data=crisis)
```

```
Warning:  non-integer #successes in a binomial glm!
```

```
plot(spanning ~ p)
lines(crisis$p,logistic.out$fitted,  col="red")
title(main="Percent spanning at critical value p")
```


Percent spanning at critical value p
spanning
0.0
0.2
0.4
0.6
0.8
1.0
0.36
0.38
0.40
0.42
0.44
0.46
p

2.23 NetLogo Code

The changes to the code are annotated below.

```
; Copyright (c) 2011-13 David O'Sullivan and George Perry
; Licensed under the Creative Commons
; Attribution-NonCommercial-ShareAlike 3.0 License
; See Info tab for full copyright and license information
;;
; Three changes for Schwitz thesis
; 1) the lattice is filled from random >= p
; 2) thus p is changed to 1-p.
; 3) pf in prop-spanning is set to 1 if the clusters span.

extensions[r profiler]

globals [
  cluster-size        ;; list of size of each occupied cluster,
                      ;; length is cluster-count - 1
  log-cluster-size    ;; log cluster size
  cluster-count       ;; total number of occupied clusters
  spanning-present?   ;; flag to denote presence of spanning cluster
  mean-size           ;; mean size of occupied, non-spanning cluster
]

patches-own [
  cluster-id ;; cluster ID assigned during tagging
  occupied?  ;; the percolation process outcome
  spanning?  ;; patch part of a spanning cluster?
  largest?  ;; patch part of largest cluster?

  r-num-code   ;; code for export to R
  ;; (0 = not occupied, 1 = occupied + !spanning, 1 = occupied + spanning)
]


;; initialise the lattice based on percolation threshold
;; r:setPlotDevice WRECKS BEHAVIORSPACE
to setup
  clear-all
;;  r:setPlotDevice

  ask patches [
    set occupied? false
    set spanning? false
```

```
    set largest? false
    set cluster-id -1
    set r-num-code 0

    set pcolor black
    if random-float 1 >= p [
      set occupied? true
      set pcolor white
      set r-num-code 1
    ]
  ]
  ;; initialise the globals
  set cluster-size []
  reset-ticks
end

to tag
  set spanning-present? false

  identify-clusters
  tag-largest-cluster
  set mean-size typical-cluster-size

  set log-cluster-size map [log ? 10] cluster-size ;-occupied
  histogram log-cluster-size
end


to colour-largest
  ask patches with [largest?] [set pcolor red]
end


;; Colour spanning cluster green
to colour-spanning
  ifelse spanning-present? [
    ask patches with [spanning?] [set pcolor green]
  ]
  [
    user-message("No spanning clusters identified on the lattice.")
  ]
end

to tag-largest-cluster
  ifelse cluster-count = 1 [
```

```
    if p = 1 [ask patches [set largest? true]]
  ]
  [
    let biggestClusterId (position (max cluster-size) cluster-size )
    ask patches with [cluster-id = (biggestClusterId)] [set largest? true]
  ]
end


to identify-clusters
  set cluster-count 0
  let occupied patches with [occupied?]
  let all-to-tag sort occupied with [cluster-id = -1]
  while [ length all-to-tag > 0 ] [
    set cluster-count cluster-count + 1
    ;; keep track of the current cluster for efficient testing if it is spanning at end
    let current-cluster patch-set nobody
    ;; lists are mutable so we use those here, starting with
    ;; a randomly selected untagged patch
    let patches-to-tag (patch-set one-of all-to-tag)

    let col random 140
    while [ any? patches-to-tag ] [
      set current-cluster (patch-set current-cluster patches-to-tag)
      ask patches-to-tag [
        ;; tag and assign cluster ID
        set cluster-id cluster-count - 1 ;; 0 will be the first cluster ID
        set pcolor col  ;; this allows user to see progress
      ]
      set patches-to-tag (patch-set [neighbors4] of patches-to-tag) with
      [occupied? and cluster-id = -1]
    ]
    let n count current-cluster
    set cluster-size lput n cluster-size
    ;; check here whether it is a spanning cluster
    if n >= min(list world-width world-height) [
      if width current-cluster = world-width or height current-cluster =
        world-height [ask current-cluster [set spanning? true]
        set spanning-present? true
      ]
    ]
    set all-to-tag filter [[cluster-id] of ? = -1] all-to-tag
    tick
  ]
  ;; restore the colours
```

```
  ask patches with [occupied?] [ set pcolor white ]
end

;; reporters for width and height of a patch-set
to-report width [p-set]
  report max [pxcor] of p-set - min [pxcor] of p-set + 1
end

to-report height [p-set]
  report max [pycor] of p-set - min [pycor] of p-set + 1
end


to-report prop-spanning
  let pf 0
  if spanning-present? [
    set pf 1
  ]
  report pf
end


to-report typical-cluster-size
  ;; mean cluster size is calcualted without the spanning cluster
  ;; and without bkgd and is size weighted
  ;; basically the typical size cluster that a rnd selected site will belong
  ;; so make a list of the cluster-size of each occupied,
  ;; non-spanning cluster patch one entry per patch
  let weighted-list map [item ([cluster-id] of ?) cluster-size]
  sort (patches with [occupied? and not spanning?])
  ;; mean of this list is the required weighted cluster size mean
  report mean weighted-list
end


;; send data to R to make log-log size rank plot
to r-hist-cs
  let s reverse sort cluster-size
  r:put "s" s

  let cs map [count-gte s ?] remove-duplicates s
  r:put "cs" cs

  r:eval("s <- unique(s)")
  r:eval("plot(x = s, y = cs, log = 'xy', las = 1, bty = 'n', xlab =
```

```
  'Clusters of size s', ylab = 'No. clusters size > s')")
end

;; reports the number of items in list lst that are >= x
to-report count-gte [lst x]
  report length filter [? >= x] lst
end

;; plot lattice in R
to lattice-to-R
  r:put "nr" world-height
  r:put "nc" world-width

  r:put "z" map [[occupied?] of ?] sort patches
  r:eval("z <-matrix(z, nrow=nr, ncol=nc)")
  r:eval("image(z, col = c('black', 'white'), asp = 1)")
end

;; send the largest cluster to R
to largest-to-r
  code-for-r

  r:put "nr" world-height
  r:put "nc" world-width

  r:put "z" map [[r-num-code] of ?] sort patches
  r:eval("z <-matrix(z, nrow=nr, ncol=nc)")
  r:eval("image(z, col = c('black', 'white', 'red'), asp = 1)")
end

;; set the r-num-code
to code-for-r
  ask patches [
    if occupied? = false [ set r-num-code 0]
    if occupied? = true and largest? = false [set r-num-code 1]
    if occupied? = true and largest? = true [set r-num-code 2]
  ]
end


to profile
  setup                  ;; set up the model
  profiler:start         ;; start profiling
  tag                    ;; run something you want to measure
  profiler:stop          ;; stop profiling
```

```
  print profiler:report  ;; view the results
  profiler:reset         ;; clear the data
end
```