

MSD 2019 Final Project

A replication and extension of Ethnicity, Insurgency, and Civil War by James D. Fearon & David D. Laitin, American Political Science Review

Preston Bradham (PMB2164), Chaim Eisenbach (CE2388), Aysha Khan (ASK2256)

2019-05-08 16:25:25

Contents

Paper Overview	1
Figures	2
Overview of Data	3
Conclusions	3
Our conclusion for the replication of Figures 1&2 (AYSHA)	3
Our conclusion for the replication of first column of table 1 (CHAIM)	3
Our conclusion for the extensions	3
- by content -	3
- predictability of model -	4
Figure 1 Replication	4
Figure 2 Replication: Part A	5
Figure 2 Replication: Part B	6
Extensions	8

Paper Overview

Fearon and Laitin analyzed post-WWII global civil wars to see if they could create a model to predict the likelihood of civil war onset. They investigated unique characteristics of the countries at hand to see if they could hypothesize which variable played the biggest role in the making of the civil war. Such as poverty, political instability, ethnic and religious diversity.

They defined civil war as conflicts that meet three criteria: involved fighting between agents of a state and organized nonstate groups who sought to take control of a government policies, the conflict killed at least 1,000 over its course with at least an average of 100 yearly deaths, and at least 100 were killed on both sides, including civilians attacked by rebels

Eleven hypotheses were investigated. For example one was: “measures of country’s ethnic or religious diversity should be associated with a higher risk of civil war.” And most of the other followed suit but with different parameters in the hypothesis.

The paper concluded that there was not enough evidence to conclude that any of the hypothesis were on target and finished with saying that civil wars are incredible hard to predict, but it is easier to predict insurgencies. It includes a cast of doubt on three wide-held notions concerning political conflict findings:

FIGURE 2. Probability of Civil War Onset per Five-Year Period

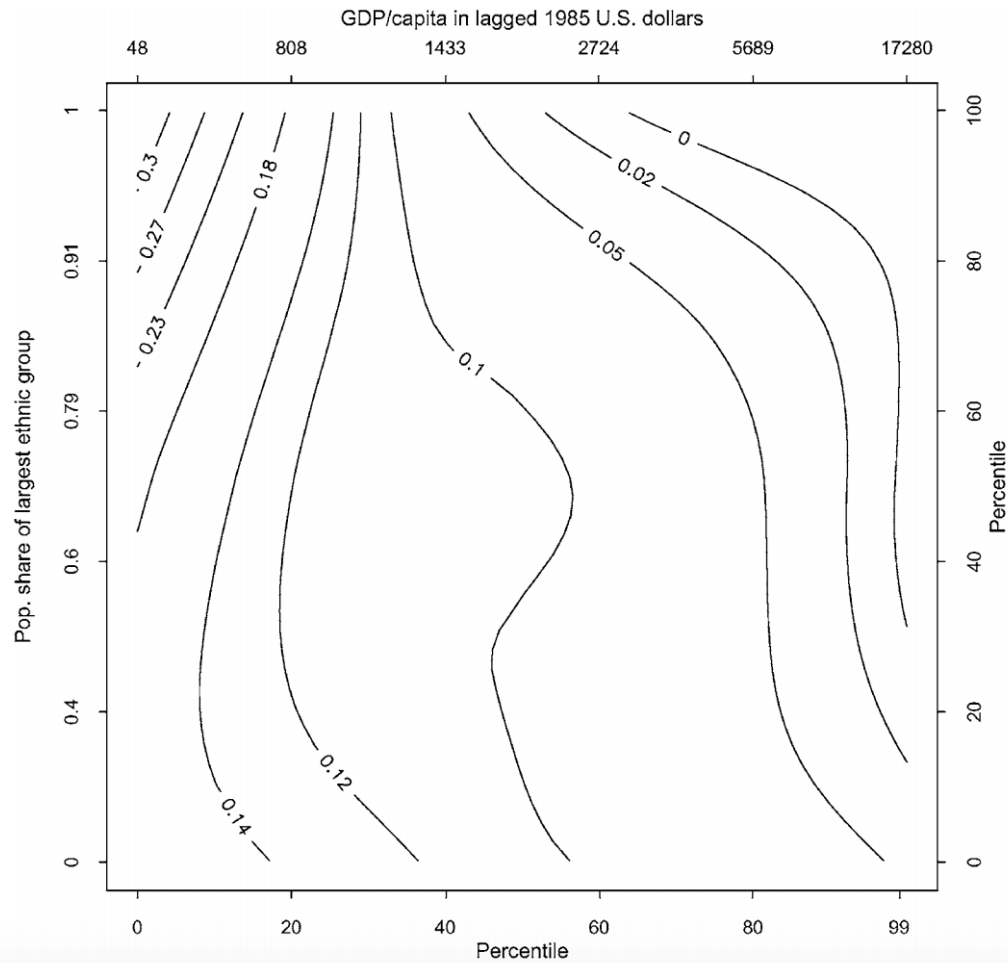


Figure 1: Figure 1

prevalence of civil war in the 1990s was not due to the end of the Cold War, greater religious and/or ethnic diversity, on its own does not make a country more prone to civil war, and cannot predict where a civil war will break out - based off of strong ethnic or political grievances.

Figures

We seek to replicate the following two figures from the paper. Figure 1 plots the number of countries with ongoing civil wars by year, from 1945 to 1999 (solid line). The paper also shows the proportion of countries with at least one ongoing war in each year (dashed line). What is interesting is that this graph indicates that post-1990s civil wars were not due to the effects of the fall of the Berlin Wall (which signified the end of the Cold War). However, conflicts associated with the fall of the Soviet Union were partly responsible for the sharp increase we witness in the early 1990s.

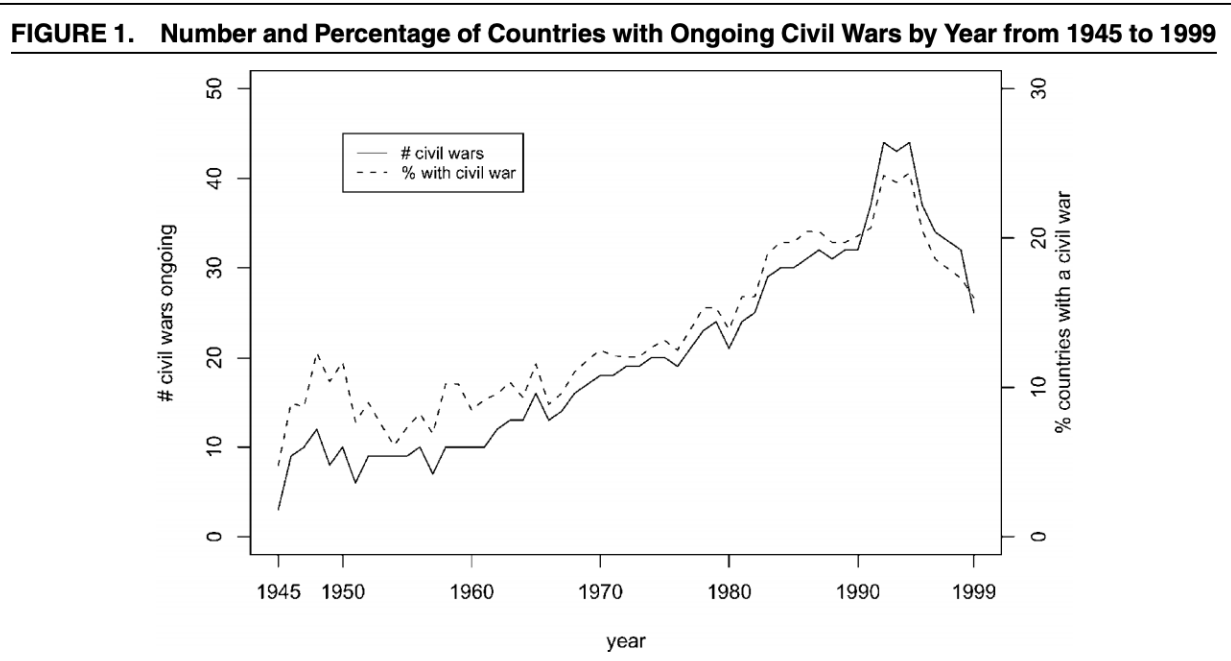


Figure 2: Figure 2

Overview of Data

Their dataset uses data across the world from the period of 1945-1999 on 161 countries that had a population of at least half a million in 1990.

It includes information on the countries: economy, location, population, employment, minerals/resources, civil war information (time frame, deaths, leader, etc.), ethnic onset, oil, GDP, Colonial country, religion percentage, and more. It allows us to explore various parameters.

Conclusions

Our conclusion for the replication of Figures 1&2 (AYSHA)

Our conclusion for the replication of first column of table 1 (CHAIM)

Our conclusion for the extensions

- by continent -

The reason we wanted to see how the civil wars broke down by continent is that we could more accurately use a historical events timeline. So we know that the end of a Cold War was not as important an event as people originally claim. The Cold War ended in 1991, and you can see from the figure it was in fact South Africa region that has an increase along with Asia, but eastern European actually saw a decrease after that time period. Before 1960 the civil wars were dominated by Asia and Eastern Europe (which historically is accurate). Then the colonial civil wars started to take place (North Africa and Middle East/ South Africa), and consistently played a key role in the overall average of the civil wars in the world.

- predictability of model -

Figure 1 Replication

```
repdata <- read.dta("./data/repdata.dta")

glance(repdata)

## Warning: 'glance.data.frame' is deprecated.
## See help("Deprecated")

## # A tibble: 1 x 4
##   nrow ncol complete.obs na.fraction
##   <int> <int>         <int>         <dbl>
## 1  6610    69           69         0.0765

# View(repdata)

sumwars_per_year <- repdata %>%
  group_by(year) %>%
  filter(war == 1) %>%
  summarize(
    count_wars_total = sum(wars)
  )

wars_per_year <- repdata %>%
  group_by(year) %>%
  filter(war == 1) %>%
  summarize(
    count_wars = sum(war)
  )

raw_num_countries <- repdata %>%
  group_by(year) %>%
  summarize(
    count_countries = sum(n())
  ) %>%
  ungroup(year)

perc_civil_war <- merge(wars_per_year, raw_num_countries, by = "year")
perc_civil_war <- merge(perc_civil_war, sumwars_per_year, by = "year")

# View(perc_civil_war)

perc_civil_war$perc <- (perc_civil_war$count_wars/perc_civil_war$count_countries)*100

plot(perc_civil_war$year, perc_civil_war$count_wars_total, axes = FALSE,
     ylim = c(0, 50), xlab = "", ylab = "", type = "l",
     col = "blue", main = "# and % of Countries with Ongoing Civil Wars (1945-1999)")
axis(2, ylim = c(0, 50), col = "black", las = 1)
mtext("# Ongoing Wars", side = 2, col = "black", line = 2.5)

# Plot the second plot and draw the axis on the right
par(new = TRUE)
```

```

plot(perc_civil_war$year, perc_civil_war$perc, pch = "solid", xlab = "", ylab = "", ylim = c(0, 30), ax
mtext("% Countries with a Civil War", side = 4, col = "black", line = 2.5)
axis(4, ylim = c(0, 30), col = "black", col.axis = "black")

# Draw the time axis
axis(1, pretty(range(perc_civil_war$year), 4))
mtext("Year", side = 1, col = "black", line = 2.5)

# Draw the legend
legend("topleft", legend = c("# civil wars", "% with civil war"),
      text.col = c("blue", "red"), col = c("blue", "red"), lty = 1:2, cex = 0.7)

```

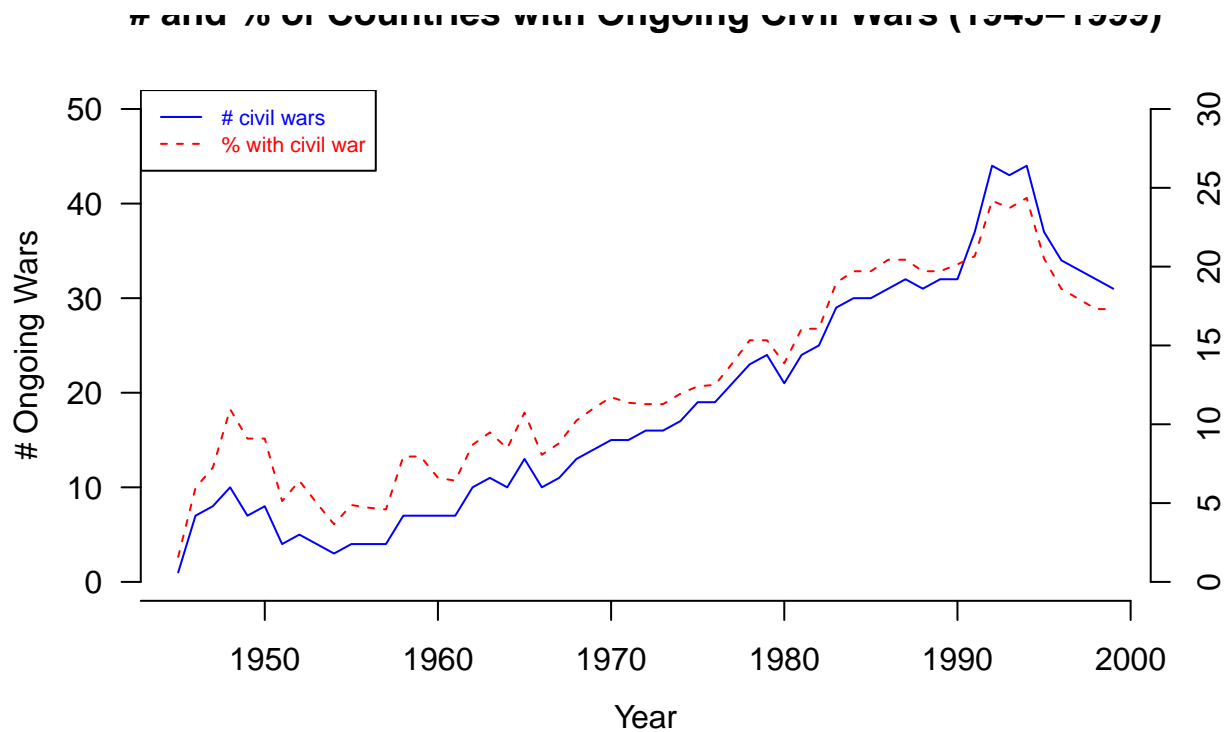


Figure 2 Replication: Part A

```

gdp_per_year <- repdata %>%
  drop_na(gdpen) %>%
  drop_na(pop) %>%
  group_by(year) %>%
  summarize(
    gdp_sum = sum(gdpen),
    pop_sum = sum(pop)
  )

gdp_per_year$gdp_pc <- gdp_per_year$gdp_sum / gdp_per_year$pop_sum * 10000

gdp_per_year_perc <- merge(gdp_per_year, perc_civil_war, by = "year")
gdp_per_year_perc$cv_percentile <- round(gdp_per_year_perc$perc / max(gdp_per_year_perc$perc), digits =

```

```
# View(gdp_per_year_perc)
```

```
gdp_per_year_perc %>%  
  ggplot(aes(x = perc, y = gdp_pc)) +  
  geom_line(color = "red")
```

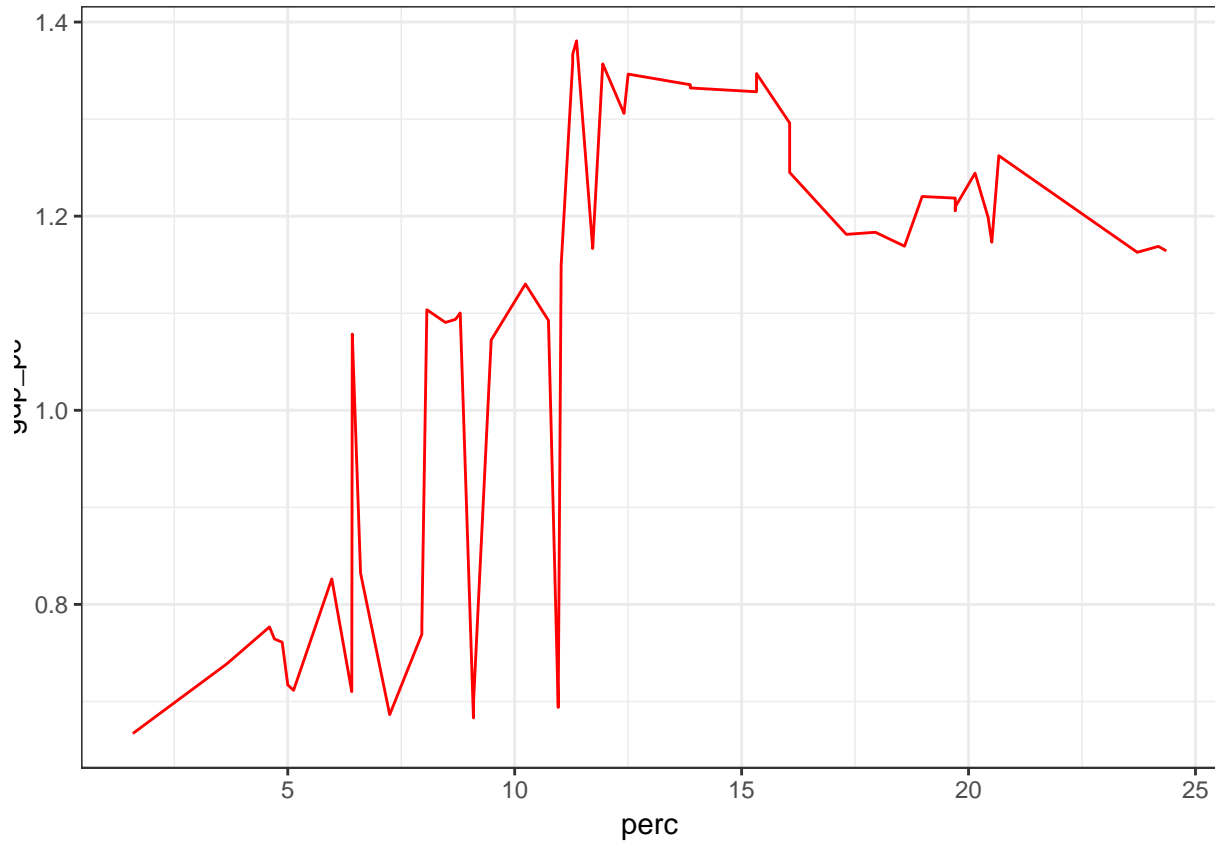


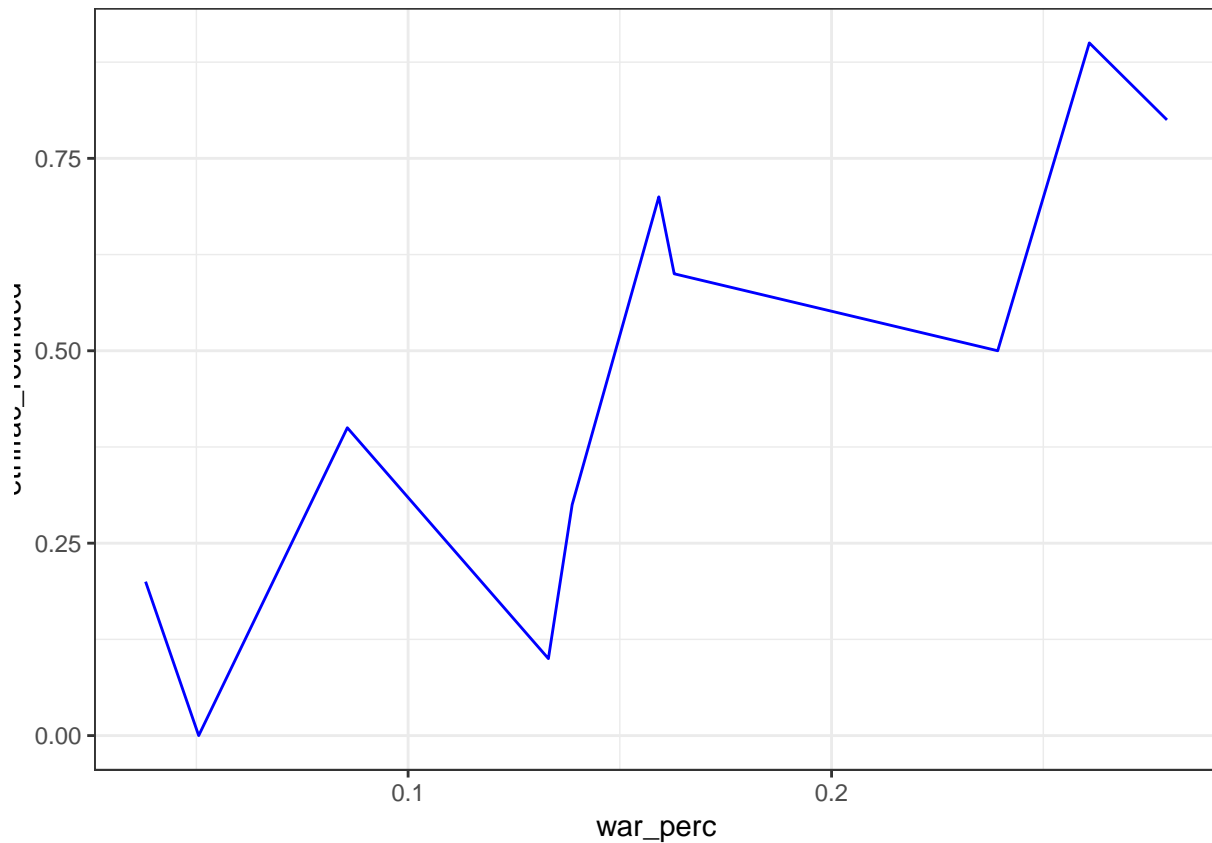
Figure 2 Replication: Part B

```
rep_data_eth <- repdata %>%  
  select(ethfrac, war) %>%  
  mutate(ethfrac_rounded = round(ethfrac, digits = 1)) %>%  
  group_by(ethfrac_rounded) %>%  
  summarize(  
    sum_countries = sum(n()),  
    sumwars_per_eth = sum(war)  
  )
```

```
rep_data_eth$war_perc = rep_data_eth$sumwars_per_eth/rep_data_eth$sum_countries
```

```
# View(rep_data_eth)
```

```
rep_data_eth %>%  
  ggplot(aes(x = war_perc, y = ethfrac_rounded)) +  
  geom_line(color = "blue")
```



```
mylogit <- glm(as.factor(onset) ~ warl + gdpenl + lpopl1 + lmtnest
+ ncontig + Oil + nwstate + instab + polity2l + ethfrac + relfrac, data = repdata, family = "binomial")
summary(mylogit)
```

```
##
## Call:
## glm(formula = as.factor(onset) ~ warl + gdpenl + lpopl1 + lmtnest +
##      ncontig + Oil + nwstate + instab + polity2l + ethfrac + relfrac,
##      family = "binomial", data = repdata)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.1412  -0.2001  -0.1446  -0.1011   3.4209
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.73134    0.73577  -9.149  < 2e-16 ***
## warl          -0.95437    0.31439  -3.036  0.002400 **
## gdpenl        -0.34386    0.07179  -4.790  1.67e-06 ***
## lpopl1         0.26294    0.07271   3.616  0.000299 ***
## lmtnest        0.21880    0.08475   2.582  0.009830 **
## ncontig        0.44333    0.27399   1.618  0.105646
## Oil           0.85764    0.27919   3.072  0.002127 **
## nwstate       1.70946    0.33859   5.049  4.45e-07 ***
## instab         0.61754    0.23510   2.627  0.008621 **
## polity2l       0.02086    0.01677   1.244  0.213536
## ethfrac        0.16641    0.37310   0.446  0.655581
```

```
## relfrac      0.28505    0.50874    0.560 0.575273
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 1077.1  on 6326  degrees of freedom
## Residual deviance:  960.8  on 6315  degrees of freedom
## (283 observations deleted due to missingness)
## AIC: 984.8
##
## Number of Fisher Scoring iterations: 8
```

Extensions

```
sumwars_per_year <- repdata %>% group_by(year) %>% filter(war == 1) %>% summarize(count_wars_total = sum(war))

wars_per_year <- repdata %>% group_by(year) %>% filter(war == 1) %>% summarize(count_wars = sum(war))

raw_num_countries <- repdata %>% group_by(year) %>% summarize(count_countries = sum(n())) %>% ungroup(year)

perc_civil_war <- merge(wars_per_year, raw_num_countries, by = "year")
perc_civil_war <- merge(perc_civil_war, sumwars_per_year, by = "year")
#perc_civil_war$perc <- (perc_civil_war$count_wars/perc_civil_war$count_countries)*100

southamerica <- repdata %>%
  filter (lamerica == 1)

wars_per_year_SA <- southamerica %>% group_by(year) %>%
  summarize(count_wars = sum(war))

percent_SA <- (wars_per_year_SA$count_wars/perc_civil_war$count_wars)
plot(wars_per_year_SA$year, percent_SA, axes = FALSE,
     ylim = c(0, 1), xlim = c(1945, 2000), xlab = "", ylab = "", type = "l",
     col = "purple", main = "% of Civil Wars by Continent (1945-1999)")

western <- repdata %>% filter (western == 1)
wars_per_year_WS <- western %>% group_by(year) %>% summarize(count_wars = sum(war))
percent_WS <- (wars_per_year_WS$count_wars/perc_civil_war$count_wars)
lines(wars_per_year_WS$year, percent_WS, col = "green")

easteurope <- repdata %>% filter (eeurop == 1)
wars_per_year_EE <- easteurope %>% group_by(year) %>% summarize(count_wars = sum(war))
percent_EE <- (wars_per_year_SA$count_wars/perc_civil_war$count_wars)
lines(wars_per_year_EE$year, percent_EE, col = "coral")

southafrica <- repdata %>% filter (ssafrica == 1)
wars_per_year_SAF <- southafrica %>% group_by(year) %>% summarize(count_wars = sum(war))
percent_SAF <- (wars_per_year_SAF$count_wars/perc_civil_war$count_wars)
lines(wars_per_year_SAF$year, percent_SAF, col = "blue")

asia <- repdata %>% filter (asia == 1)
wars_per_year_AS <- asia %>% group_by(year) %>% summarize(count_wars = sum(war))
```



```

percent_AS <- (wars_per_year_AS$count_wars/perc_civil_war$count_wars)
lines(wars_per_year_AS$year, percent_AS, col = "red")

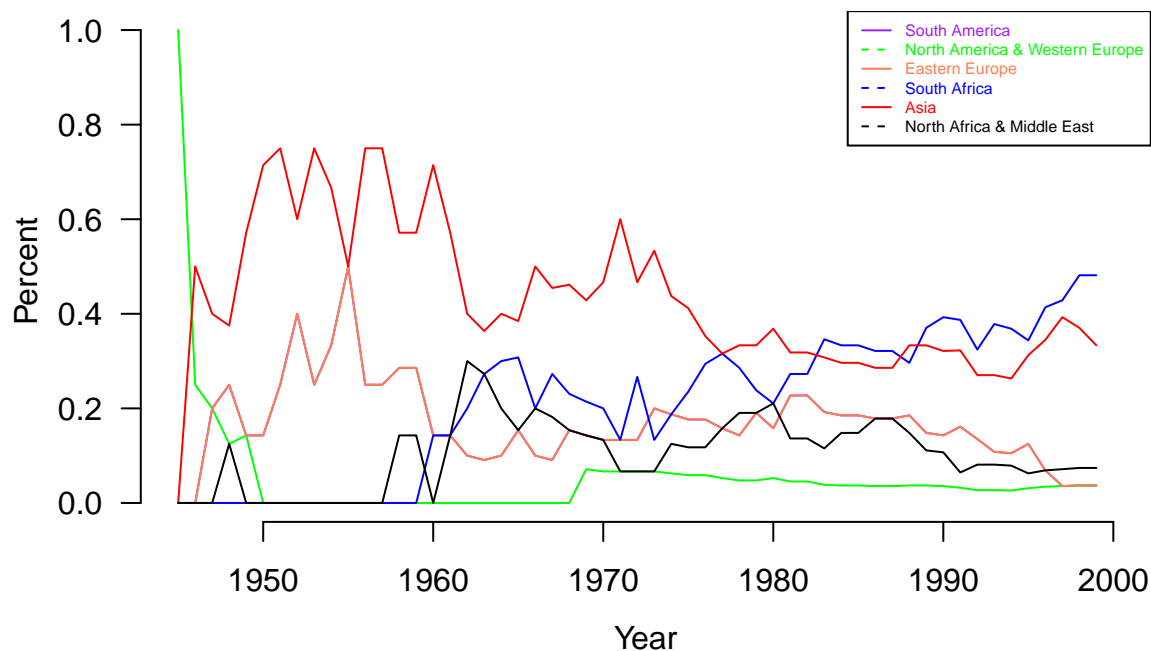
northafricamiddleeast <- repdata %>% filter (nafrme == 1)
wars_per_year_NAM <- northafricamiddleeast %>% group_by(year) %>% summarize(count_wars = sum(war))
percent_NAM <- (wars_per_year_NAM$count_wars/perc_civil_war$count_wars)
lines(wars_per_year_NAM$year, percent_NAM, col = "black")

#sumwars_per_year <- repdata %>% group_by(year) %>% filter(war == 1) %>% summarize(count_wars = sum(war))
#lines(sumwars_per_year$year, sumwars_per_year$count_wars, col = "darkgreen")

axis(2, ylim = c(0, 1), col = "black", las = 1)
axis(1, xlim = c(1945, 1999), col = "black", las = 1)
mtext("Percent", side = 2, col = "black", line = 2.5)
mtext("Year", side = 1, col = "black", line = 2.5)
#legend("topleft", legend = c("South America", "North America & Western Europe", "Eastern Europe", "South Africa", "Asia", "North Africa & Middle East"),
#text.col = c("purple", "green", "coral", "blue", "red", "orange", "darkgreen"), col = c("purple", "green", "coral", "blue", "red", "orange", "darkgreen"))
legend("topright", legend = c("South America", "North America & Western Europe", "Eastern Europe", "South Africa", "Asia", "North Africa & Middle East"),
text.col = c("purple", "green", "coral", "blue", "red", "black", "darkgreen"), col = c("purple", "green", "coral", "blue", "red", "black", "darkgreen"))

```

% of Civil Wars by Continent (1945-1999)



```

repdata <- repdata %>% group_by(war) %>% mutate(outcome = ifelse(war == 1, 'civil_war', 'no_war'))
repdata$outcome <- as.factor(repdata$outcome)
# View(repdata)

set.seed(42)
ndx <- sample(nrow(repdata), floor(nrow(repdata) * 0.9))
train <- repdata[ndx,]
test <- repdata[-ndx,]

xTrain <- train[, -70]

```

```

yTrain <- train$outcome

xTest <- test[, -70]
yTest <- test$outcome

# model <- naiveBayes(xTrain, yTrain)
# summary(model)

model <- naiveBayes(outcome ~ war1 + gdpen1 + lpopl1 + lmtnest
+ ncontig + Oil + nwstate + instab + polity21 + ethfrac + relfrac, data = train, family = "binomial")
summary(model)

##           Length Class  Mode
## apriori      2      table numeric
## tables      11      -none- list
## levels       2      -none- character
## isnumeric    11      -none- logical
## call         5      -none- call

df <- data.frame(actual = yTest, pred = predict(model, test))

## Warning in data.matrix(newdata): NAs introduced by coercion
## Warning in data.matrix(newdata): NAs introduced by coercion
## Warning in data.matrix(newdata): NAs introduced by coercion
## Warning in data.matrix(newdata): NAs introduced by coercion

head(df)

##   actual   pred
## 1 no_war no_war
## 2 no_war no_war
## 3 no_war no_war
## 4 no_war no_war
## 5 no_war no_war
## 6 no_war no_war

table(df)

##           pred
## actual      civil_war no_war
##   civil_war         90      9
##   no_war           5    557

# accuracy: fraction of correct classifications
df %>%
  summarize(acc = mean(pred == actual))

##           acc
## 1 0.97882

# precision: fraction of positive predictions that are actually true
df %>%
  filter(pred == 'civil_war') %>%
  summarize(prec = mean(actual == 'civil_war'))

```

```

##           prec
## 1 0.9473684
# recall: fraction of true examples that we predicted to be positive
# aka true positive rate, sensitivity
df %>%
  filter(actual == 'civil_war') %>%
  summarize(recall = mean(pred == 'civil_war'))

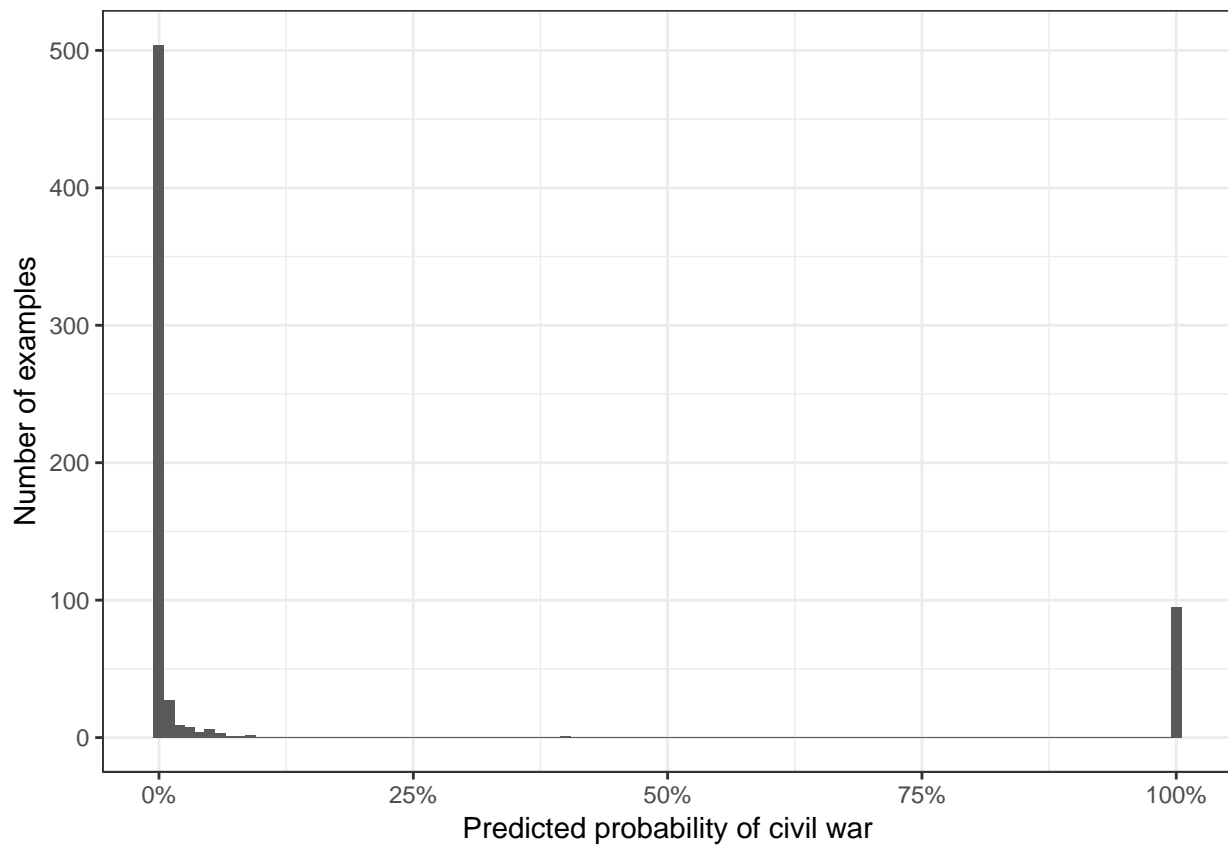
##           recall
## 1 0.9090909
# false positive rate: fraction of false examples that we predicted to be positive
df %>%
  filter(actual == 'no_war') %>%
  summarize(fpr = mean(pred == 'civil_war'))

##           fpr
## 1 0.008896797
# plot histogram of predicted probabilities
# note overconfident predictions
probs <- data.frame(predict(model, test, type="raw"))

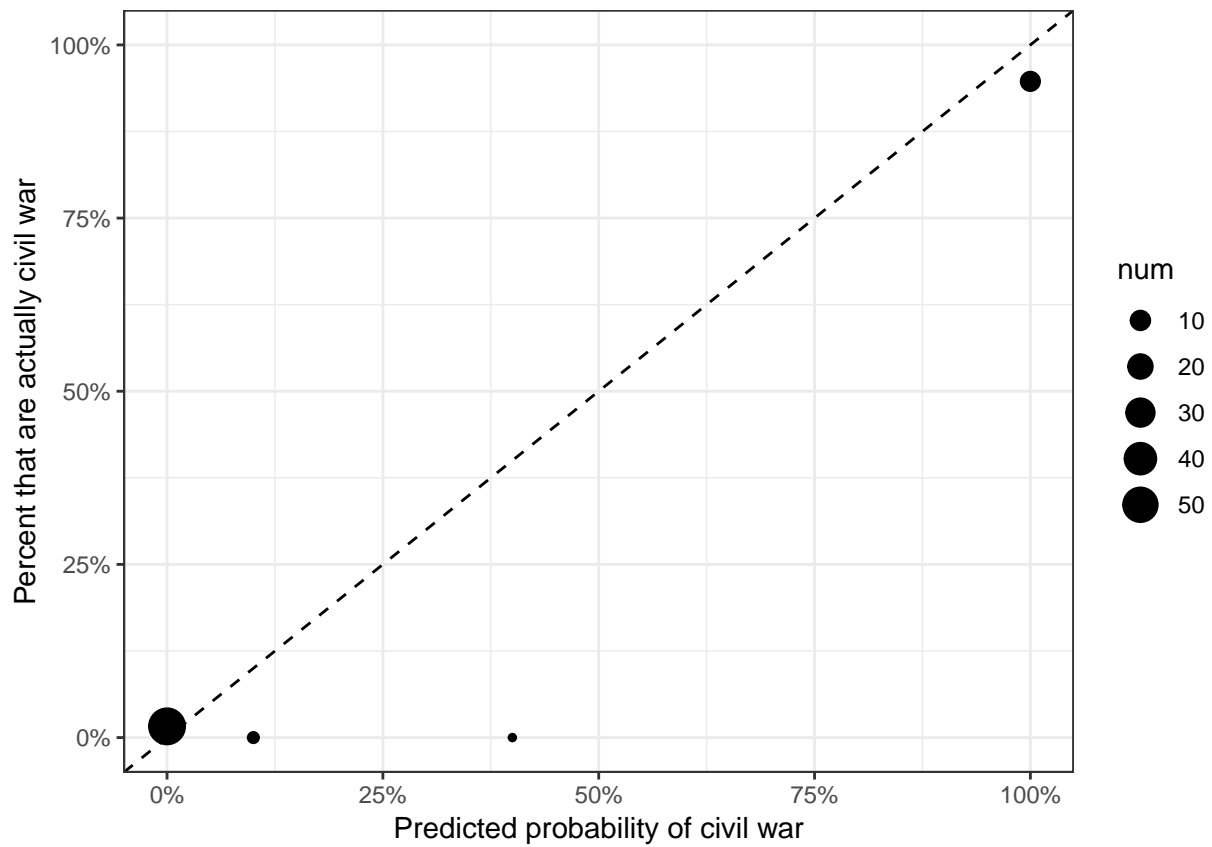
## Warning in data.matrix(newdata): NAs introduced by coercion
## Warning in data.matrix(newdata): NAs introduced by coercion
## Warning in data.matrix(newdata): NAs introduced by coercion
## Warning in data.matrix(newdata): NAs introduced by coercion

ggplot(probs, aes(x = civil_war)) +
  geom_histogram(binwidth = 0.01) +
  scale_x_continuous(label = percent) +
  xlab('Predicted probability of civil war') +
  ylab('Number of examples')

```



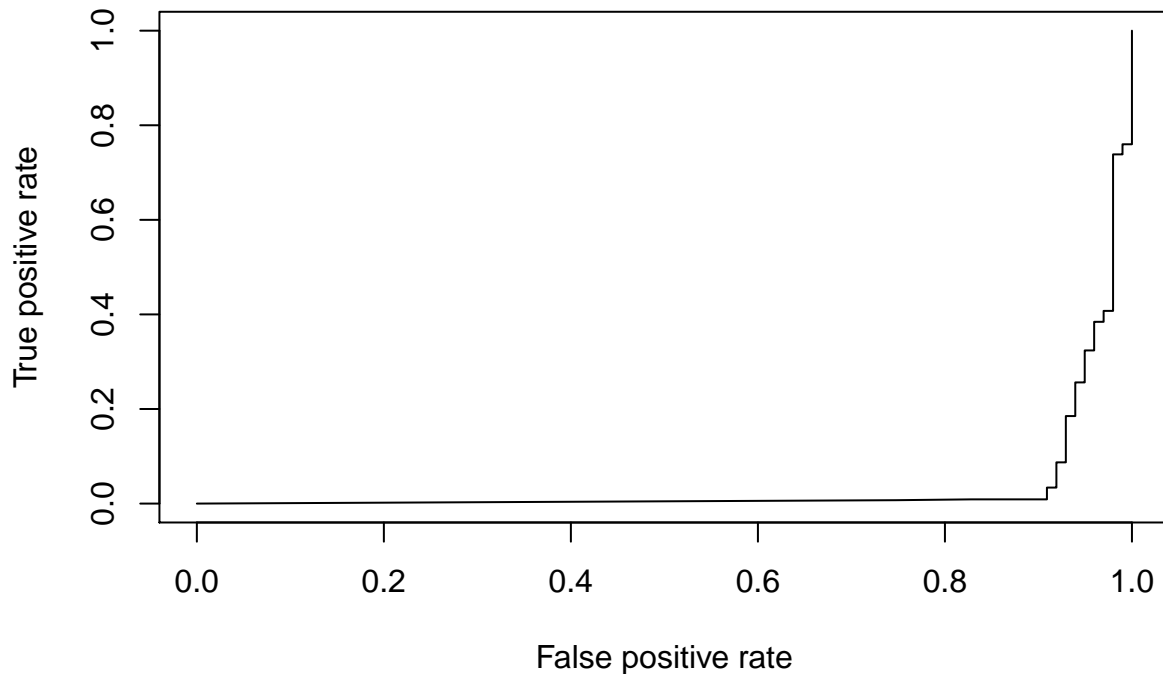
```
data.frame(predicted=probs[, "civil_war"], actual=yTest) %>%
  group_by(predicted=round(predicted*10)/10) %>%
  summarize(num=n(), actual=mean(actual == "civil_war")) %>%
  ggplot(data=., aes(x=predicted, y=actual, size=num)) +
  geom_point() +
  geom_abline(linetype=2) +
  scale_x_continuous(labels=percent, lim=c(0,1)) +
  scale_y_continuous(labels=percent, lim=c(0,1)) +
  xlab('Predicted probability of civil war') +
  ylab('Percent that are actually civil war')
```



```
# create a ROCR object
pred <- prediction(probs[, "civil_war"], yTest)

# create a ROCR object
pred <- prediction(probs[, "civil_war"], yTest)

# plot ROC curve
perf_nb <- performance(pred, measure='tpr', x.measure='fpr')
plot(perf_nb)
```



```
performance(pred, 'auc')
```

```
## An object of class "performance"
## Slot "x.name":
## [1] "None"
##
## Slot "y.name":
## [1] "Area under the ROC curve"
##
## Slot "alpha.name":
## [1] "none"
##
## Slot "x.values":
## list()
##
## Slot "y.values":
## [[1]]
## [1] 0.03610842
##
##
## Slot "alpha.values":
## list()
```

```
sessionInfo()
```

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-apple-darwin15.6.0 (64-bit)
## Running under: macOS 10.14.4
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.5/Resources/lib/libRlapack.dylib
##
```

```

## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] grid      stats      graphics  grDevices utils      datasets  methods
## [8] base
##
## other attached packages:
## [1] e1071_1.7-1      pROC_1.14.0      ROCR_1.0-7      gplots_3.0.1.1
## [5] scales_1.0.0     broom_0.5.2      lubridate_1.7.4 forcats_0.4.0
## [9] stringr_1.4.0    dplyr_0.8.0.1    purrr_0.3.2     readr_1.3.1
## [13] tidyr_0.8.3      tibble_2.1.1     ggplot2_3.1.1   tidyverse_1.2.1
## [17] foreign_0.8-70
##
## loaded via a namespace (and not attached):
## [1] gtools_3.8.1      tidyselect_0.2.5  xfun_0.6
## [4] haven_2.1.0       lattice_0.20-35   colorspace_1.4-1
## [7] generics_0.0.2    htmltools_0.3.6   yaml_2.2.0
## [10] utf8_1.1.4        rlang_0.3.4       pillar_1.3.1
## [13] glue_1.3.1        withr_2.1.2       modelr_0.1.4
## [16] readxl_1.3.1      plyr_1.8.4        munsell_0.5.0
## [19] gtable_0.3.0      cellranger_1.1.0  rvest_0.3.3
## [22] caTools_1.17.1.2  evaluate_0.13     labeling_0.3
## [25] knitr_1.22        class_7.3-14      fansi_0.4.0
## [28] Rcpp_1.0.1        KernSmooth_2.23-15 backports_1.1.4
## [31] gdata_2.18.0      jsonlite_1.6      hms_0.4.2
## [34] digest_0.6.18     stringi_1.4.3     cli_1.1.0
## [37] tools_3.5.1       bitops_1.0-6      magrittr_1.5
## [40] lazyeval_0.2.2    crayon_1.3.4      pkgconfig_2.0.2
## [43] xml2_1.2.0        assertthat_0.2.1  rmarkdown_1.12
## [46] httr_1.4.0        rstudioapi_0.10   R6_2.4.0
## [49] nlme_3.1-137      compiler_3.5.1

```