# POLI 271 Problem set 4

## POLI 271 Problem set 4

### Problem 1

**a**

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(gt)
```

```
Warning: package 'gt' was built under R version 4.4.2
```

```
library(modelsummary)
```

```
Warning: package 'modelsummary' was built under R version 4.4.2
```

```
`modelsummary` 2.0.0 now uses `tinytable` as its default table-drawing
  backend. Learn more at: https://vincentarelbundock.github.io/tinytable/

Revert to `kableExtra` for one session:

  options(modelsummary_factory_default = 'kableExtra')
  options(modelsummary_factory_latex = 'kableExtra')
  options(modelsummary_factory_html = 'kableExtra')

Silence this message forever:

  config_modelsummary(startup_message = FALSE)
```

```
library(stargazer)
```

```
Please cite as:

 Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
 R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(ROCR)
```

```
Warning: package 'ROCR' was built under R version 4.4.2
```

```
library(caret)
```

```
Warning: package 'caret' was built under R version 4.4.2

Loading required package: lattice

Attaching package: 'caret'

The following object is masked from 'package:purrr':

    lift
```

```
library(cvTools)
```

```
Warning: package 'cvTools' was built under R version 4.4.2

Loading required package: robustbase

Warning: package 'robustbase' was built under R version 4.4.2
```

```r
library(MASS)
```

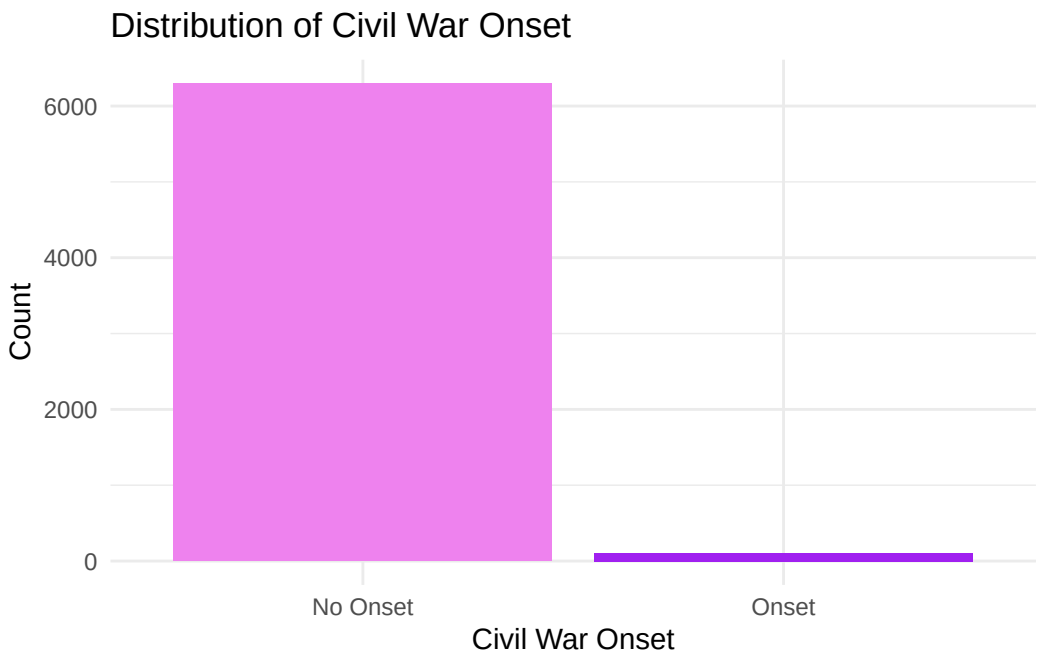
```
Attaching package: 'MASS'

The following object is masked from 'package:dplyr':

    select
```

```r
flmdw <- read.csv("flmdw-1.csv")
head(flmdw, 20)
```

```
    X country year onset instab warl gdpenl   lpopl1  lmtnest ncontig Oil
1   1     USA 1945     0      0    0  7.626 11.85630 3.214868       1   0
2   2     USA 1946     0      0    0  7.626 11.85630 3.214868       1   0
3   3     USA 1947     0      0    0  7.654 11.86313 3.214868       1   0
4   4     USA 1948     0      0    0  8.025 11.86859 3.214868       1   0
5   5     USA 1949     0      0    0  8.270 11.88673 3.214868       1   0
6   6     USA 1950     0      0    0  8.040 11.90488 3.214868       1   0
7   7     USA 1951     0      0    0  8.772 11.93343 3.214868       1   0
8   8     USA 1952     0      0    0  9.109 11.95118 3.214868       1   0
9   9     USA 1953     0      0    0  9.074 11.96862 3.214868       1   0
10 10     USA 1954     0      0    0  9.300 11.98589 3.214868       1   0
11 11     USA 1955     0      0    0  9.089 12.00274 3.214868       1   0
12 12     USA 1956     0      0    0  9.723 12.01932 3.214868       1   0
13 13     USA 1957     0      0    0  9.712 12.03696 3.214868       1   0
14 14     USA 1958     0      0    0  9.643 12.05429 3.214868       1   0
15 15     USA 1959     0      0    0  9.370 12.07121 3.214868       1   0
16 16     USA 1960     0      0    0  9.839 12.08797 3.214868       1   0
17 17     USA 1961     0      0    0  9.895 12.10444 3.214868       1   0
18 18     USA 1962     0      0    0  9.946 12.12099 3.214868       1   0
19 19     USA 1963     0      0    0 10.358 12.13638 3.214868       1   0
20 20     USA 1964     0      0    0 10.642 12.15079 3.214868       1   0
   nwstate polity2l   ethfrac relfrac
```

```
1         0        10 0.3569501   0.596
2         0        10 0.3569501   0.596
3         0        10 0.3569501   0.596
4         0        10 0.3569501   0.596
5         0        10 0.3569501   0.596
6         0        10 0.3569501   0.596
7         0        10 0.3569501   0.596
8         0        10 0.3569501   0.596
9         0        10 0.3569501   0.596
10        0        10 0.3569501   0.596
11        0        10 0.3569501   0.596
12        0        10 0.3569501   0.596
13        0        10 0.3569501   0.596
14        0        10 0.3569501   0.596
15        0        10 0.3569501   0.596
16        0        10 0.3569501   0.596
17        0        10 0.3569501   0.596
18        0        10 0.3569501   0.596
19        0        10 0.3569501   0.596
20        0        10 0.3569501   0.596
```

```r
library(ggplot2)
ggplot(flmdw, aes(x = factor(onset))) +
  geom_bar(fill = c("violet", "purple")) +
  labs(title = "Distribution of Civil War Onset",
       x = "Civil War Onset",
       y = "Count") +
  scale_x_discrete(labels = c("No Onset", "Onset")) +
  theme_minimal()
```




This is a rare event. Bayesian Logistic Regression; LASSO

**b**

```
library(dplyr)
colnames(flmdw)
```

```
 [1] "X"        "country"  "year"     "onset"    "instab"   "warl"
 [7] "gdpenl"   "lpopl1"   "lmtnest"  "ncontig"  "Oil"      "nwstate"
[13] "polity2l" "ethfrac"  "relfrac"
```

```
flmdw_complete <- flmdw %>%
  dplyr::select(onset, gdpenl, lpopl1, lmtnest, Oil, polity2l, relfrac) %>%
  na.omit()

model1 <- glm(onset ~ gdpenl + lpopl1 + lmtnest,
              data = flmdw_complete, family = binomial)
model2 <- glm(onset ~ gdpenl + lpopl1 + lmtnest,
              data = flmdw_complete, family = binomial(link = "probit"))
```

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

```r
model3 <- glm(onset ~ gdpenl + lpopl1 + lmtnest + Oil + polity2l + relfrac,
              data = flmdw_complete, family = binomial(link = "probit"))
```

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

```r
model4 <- glm(onset ~ gdpenl + lpopl1 + lmtnest + Oil + polity2l * relfrac,
              data = flmdw_complete, family = binomial(link = "probit"))
```

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

```r
summary(model1)
```

```
Call:
glm(formula = onset ~ gdpenl + lpopl1 + lmtnest, family = binomial,
    data = flmdw_complete)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -6.04197    0.61198  -9.873  < 2e-16 ***
gdpenl      -0.29518    0.06235  -4.734  2.2e-06 ***
lpopl1       0.23632    0.06169   3.831 0.000128 ***
lmtnest      0.17810    0.07996   2.227 0.025923 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1079.6  on 6401  degrees of freedom
Residual deviance: 1013.1  on 6398  degrees of freedom
AIC: 1021.1

Number of Fisher Scoring iterations: 8
```

```r
summary(model2)
```

```
Call:
glm(formula = onset ~ gdpenl + lpopl1 + lmtnest, family = binomial(link = "probit"),
```

```
    data = flmdw_complete)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.96613    0.25589 -11.591  < 2e-16 ***
gdpenl      -0.11373    0.02314  -4.914 8.92e-07 ***
lpopl1       0.09964    0.02664   3.740 0.000184 ***
lmtnest      0.07451    0.03199   2.329 0.019843 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1079.6  on 6401  degrees of freedom
Residual deviance: 1012.4  on 6398  degrees of freedom
AIC: 1020.4

Number of Fisher Scoring iterations: 8
```

```
summary(model3)
```

```
Call:
glm(formula = onset ~ gdpenl + lpopl1 + lmtnest + Oil + polity2l +
    relfrac, family = binomial(link = "probit"), data = flmdw_complete)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.950436   0.273591 -10.784  < 2e-16 ***
gdpenl      -0.139777   0.026540  -5.267 1.39e-07 ***
lpopl1       0.088005   0.026775   3.287 0.001013 **
lmtnest      0.085062   0.033273   2.556 0.010574 *
Oil          0.404288   0.116519   3.470 0.000521 ***
polity2l     0.013098   0.006503   2.014 0.043981 *
relfrac      0.222227   0.198665   1.119 0.263311
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1079.63  on 6401  degrees of freedom
Residual deviance:  999.01  on 6395  degrees of freedom
```

```
AIC: 1013

Number of Fisher Scoring iterations: 8
```

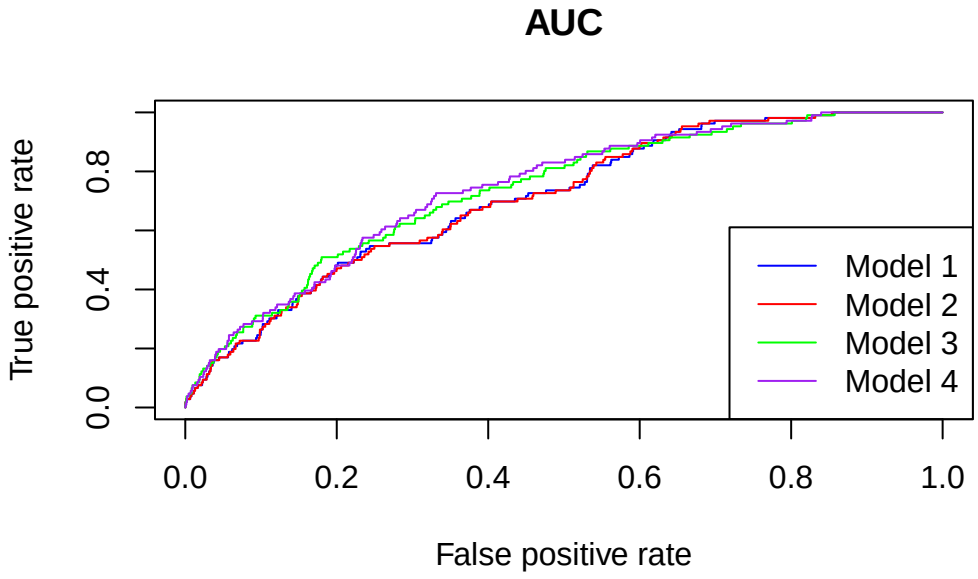
```r
summary(model4)
```

```
Call:
glm(formula = onset ~ gdpenl + lpopl1 + lmtnest + Oil + polity2l *
    relfrac, family = binomial(link = "probit"), data = flmdw_complete)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -3.037835   0.279631 -10.864  < 2e-16 ***
gdpenl           -0.147410   0.026829  -5.494 3.92e-08 ***
lpopl1            0.093681   0.026971   3.473 0.000514 ***
lmtnest           0.089834   0.033466   2.684 0.007267 **
Oil               0.402871   0.117037   3.442 0.000577 ***
polity2l         -0.006433   0.012563  -0.512 0.608617
relfrac           0.318370   0.202051   1.576 0.115097
polity2l:relfrac  0.053889   0.029070   1.854 0.063767 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1079.63  on 6401  degrees of freedom
Residual deviance:  995.66  on 6394  degrees of freedom
AIC: 1011.7

Number of Fisher Scoring iterations: 8
```

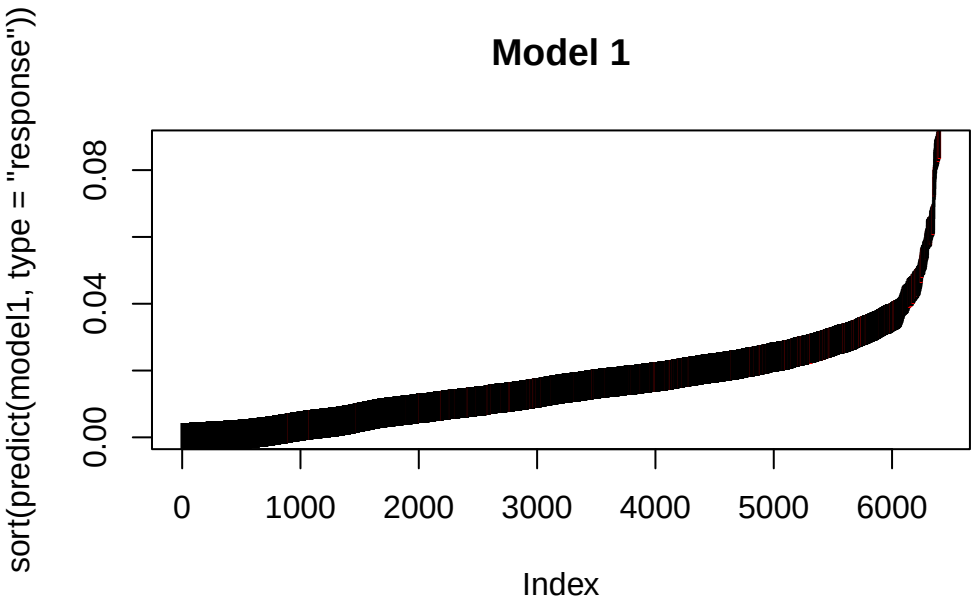
**c**

```r
pred1 <- prediction(predict(model1, type = "response"), flmdw_complete$onset)
roc1 <- performance(pred1, "tpr", "fpr")

pred2 <- prediction(predict(model2, type = "response"), flmdw_complete$onset)
roc2 <- performance(pred2, "tpr", "fpr")
```

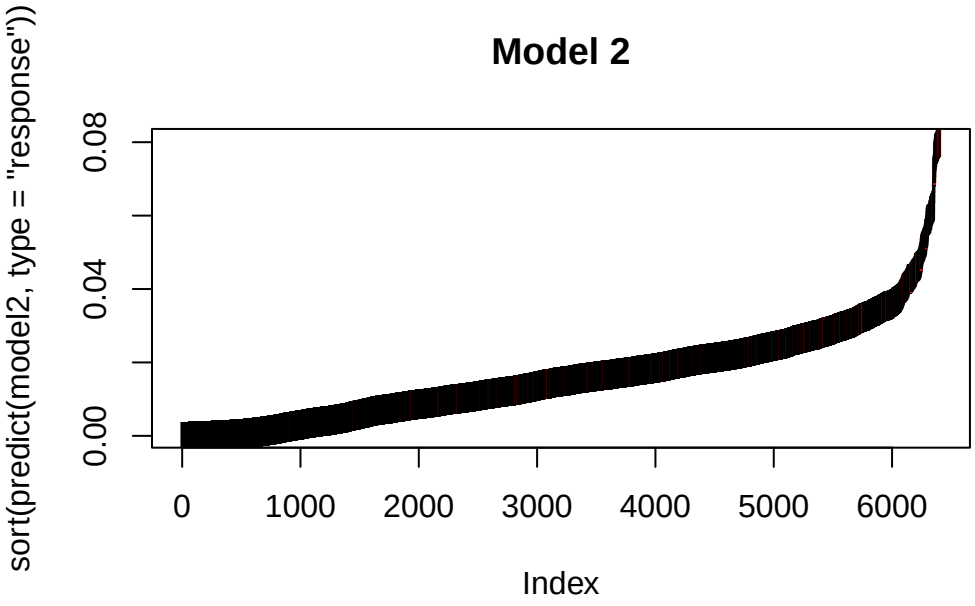
```r
pred3 <- prediction(predict(model3, type = "response"), flmdw_complete$onset)
roc3 <- performance(pred3, "tpr", "fpr")

pred4 <- prediction(predict(model4, type = "response"), flmdw_complete$onset)
roc4 <- performance(pred4, "tpr", "fpr")

plot(roc1, col = "blue", main = "AUC")
plot(roc2, col = "red", add = TRUE)
plot(roc3, col = "green", add = TRUE)
plot(roc4, col = "purple", add = TRUE)
legend("bottomright", legend = c("Model 1", "Model 2", "Model 3", "Model 4"),
       col = c("blue", "red", "green", "purple"), lty = 1)
```



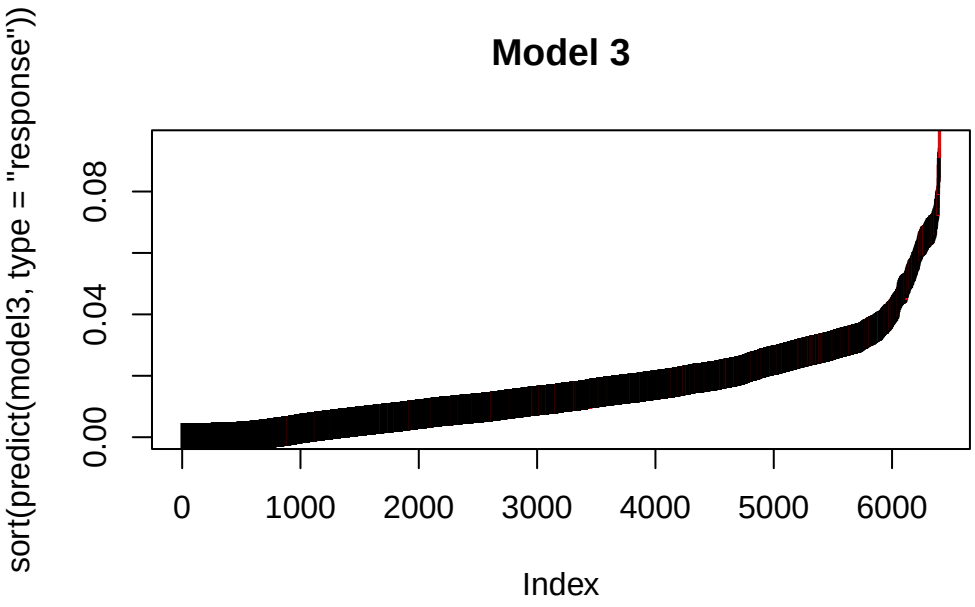

```r
plot(sort(predict(model1, type = "response")), col = ifelse(flmdw_complete$onset[order(predi
```



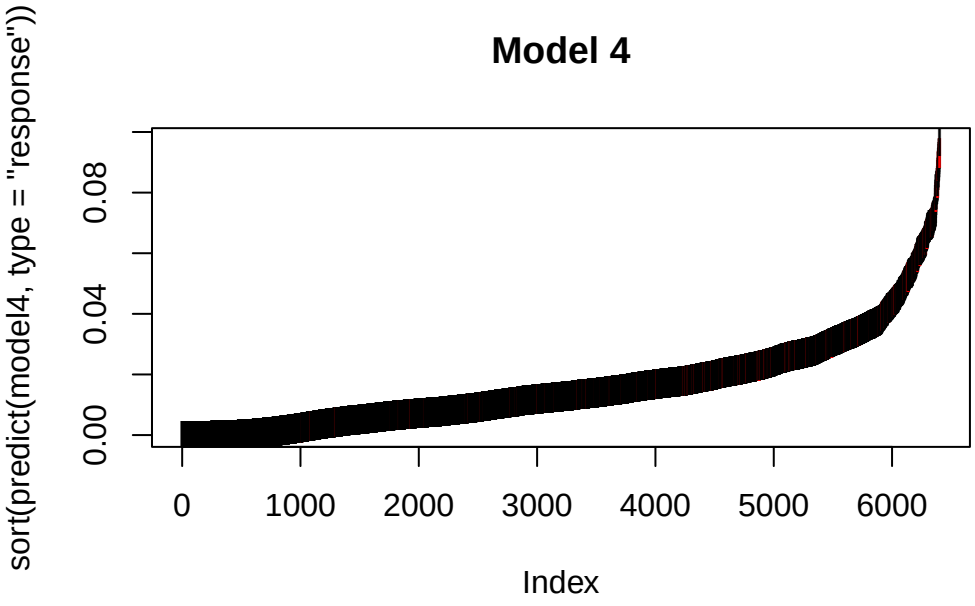

```
plot(sort(predict(model2, type = "response")), col = ifelse(flmdw_complete$onset[order(predi
```



```
plot(sort(predict(model3, type = "response")), col = ifelse(flmdw_complete$onset[order(predi
```




```
plot(sort(predict(model4, type = "response")), col = ifelse(flmdw_complete$onset[order(predi
```



**d**

```
model_restricted <- glm(onset ~ gdpenl + lpopl1 + lmtnest + Oil + relfrac,
                        data = flmdw_complete, family = binomial(link = "probit"))
```

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

```
lr_test <- anova(model_restricted, model4, test = "Chisq")

print(lr_test)
```

```
Analysis of Deviance Table

Model 1: onset ~ gdpenl + lpopl1 + lmtnest + Oil + relfrac
Model 2: onset ~ gdpenl + lpopl1 + lmtnest + Oil + polity2l * relfrac
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      6396    1003.01
2      6394     995.66  2   7.3558  0.02528 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value (0.02528) is less than 0.05, meaning we reject the null hypothesis that $\beta_{dem} = \beta_{demfrac} = 0$

**e**

```
library(caret)
library(pROC)
```

```
Warning: package 'pROC' was built under R version 4.4.2
```

```
Type 'citation("pROC")' for a citation.
```

```
Attaching package: 'pROC'
```

```
The following objects are masked from 'package:stats':

    cov, smooth, var
```

```
set.seed(3407)

flmdw_complete$onset <- factor(flmdw_complete$onset, levels = c(0,1), labels = c("No", "Yes")

cv_control <- trainControl(method = "cv", number = 10, classProbs = TRUE, summaryFunction = t

cv_model1 <- train(onset ~ gdpenl + lpopl1 + lmtnest + Oil + relfrac,
                   data = flmdw_complete, method = "glm", family = binomial,
                   trControl = cv_control, metric = "ROC")

cv_model2 <- train(onset ~ gdpenl + lpopl1 + lmtnest + Oil + polity2l * relfrac,
                   data = flmdw_complete, method = "glm", family = binomial,
                   trControl = cv_control, metric = "ROC")
```

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

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

```
prob1 <- predict(cv_model1, flmdw_complete, type = "prob")[,"Yes"]
prob2 <- predict(cv_model2, flmdw_complete, type = "prob")[,"Yes"]

roc1 <- roc(flmdw_complete$onset, prob1)
```

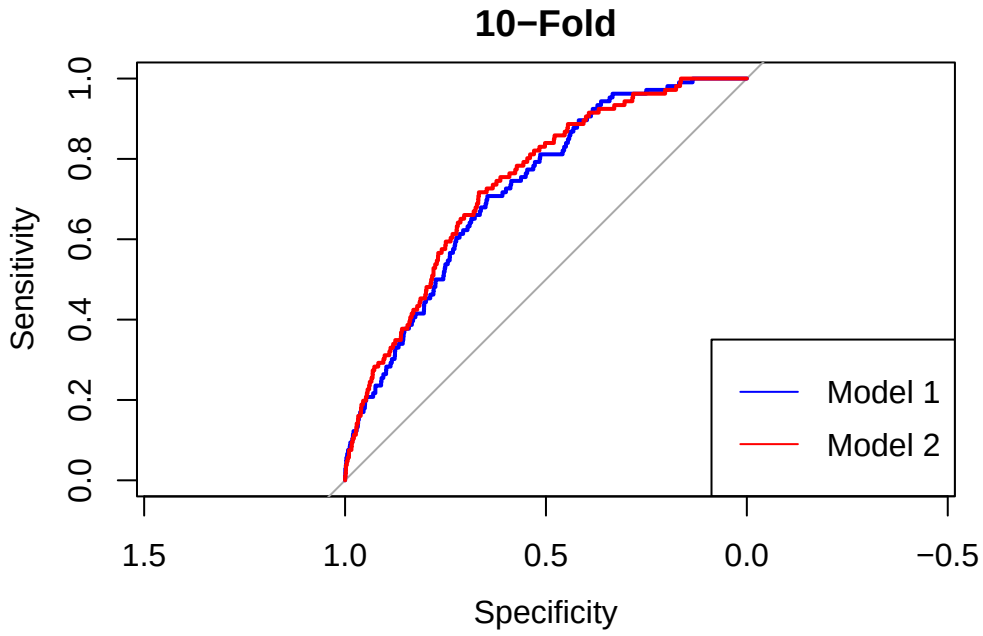
```
Setting levels: control = No, case = Yes
```

```
Setting direction: controls < cases
```

```
roc2 <- roc(flmdw_complete$onset, prob2)
```

```
Setting levels: control = No, case = Yes
Setting direction: controls < cases
```

```
plot(roc1, col = "blue", main = "10-Fold")
lines(roc2, col = "red")
legend("bottomright", legend = c("Model 1", "Model 2"), col = c("blue", "red"), lty = 1)
```




```
auc(roc1)
```

```
Area under the curve: 0.726
```

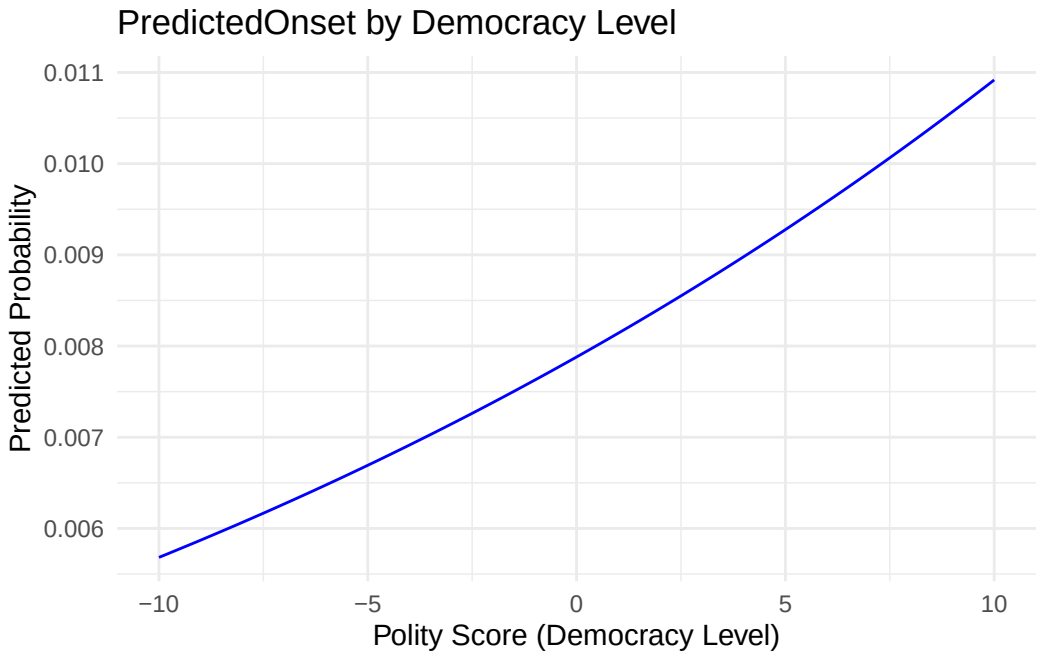
```
auc(roc2)
```

```
Area under the curve: 0.7402
```

**f**

```
prediction_data <- expand.grid(
  polity2l = seq(min(flmdw_complete$polity2l), max(flmdw_complete$polity2l), length.out = 100
  relfrac = mean(flmdw_complete$relfrac),
  gdpenl = mean(flmdw_complete$gdpenl),
  lpopl1 = mean(flmdw_complete$lpopl1),
```

```
  lmtnest = mean(flmdw_complete$lmtnest),
  Oil = 0
)

prediction_data$predicted <- predict(cv_model2, newdata = prediction_data, type = "prob")[,"

ggplot(prediction_data, aes(x = polity2l, y = predicted)) +
  geom_line(color = "blue") +
  labs(title = "PredictedOnset by Democracy Level",
       x = "Polity Score (Democracy Level)",
       y = "Predicted Probability") +
  theme_minimal()
```




```
interaction_data <- expand.grid(
  polity2l = seq(min(flmdw_complete$polity2l), max(flmdw_complete$polity2l), length.out = 100
  relfrac = c(0.1, 0.5, 0.9),
  gdpenl = mean(flmdw_complete$gdpenl),
  lpopl1 = mean(flmdw_complete$lpopl1),
  lmtnest = mean(flmdw_complete$lmtnest),
  Oil = 0
)
```

```r
interaction_data$predicted <- predict(cv_model2, newdata = interaction_data, type = "prob")[

ggplot(interaction_data, aes(x = polity2l, y = predicted, color = factor(relfrac))) +
  geom_line() +
  labs(title = "Interaction Effect of Democracy and Religious Fractionalization",
       x = "Polity Score",
       y = "Predicted Probability of Civil War Onset",
       color = "Religious Fractionalization") +
  theme_minimal()
```