Gov 2001: Problem Set 5

Due Wednesday, March 9th at 6pm

Instructions

You should submit your answers and R code to the problems below using the Quizzes section on Canvas.

Problem 1 - Replicating Fearon and Laitin (2003)

For this problem you will replicate and explore the results of the article "Ethnicity, Insurgency and Civil War," by James Fearon and David Laitin (*American Political Science Review*. 97 (1) February 2003.). This article uses a number of institutional, geographic and social variables which have been hypothesized to predict the onset of civil conflict.

The complete dataset is available on the website in the .dta format. You will use the following variables in your analysis:

- onset: civil war onset that year
- warl: a 'distinct' civil war ongoing in previous year
- gdpenl: lagged per capita income, thousands of US dollars
- lpopl1: lagged natural logarithm of population
- lmtnest: log
- ncontig: country is non-contiguous geographically
- Oil: country is an oil exporter
- nwstate: country achieved independence in past two years
- instab: significant change in polity score in last 3 years, lagged
- polity2l: lagged level of democratization
- ethfrac: a measure of ethnic fragmentation
- relfrac: a measure of religious fragmentation

Once you have loaded the data, you will need to subset it to include only the above variables. Then delete all rows containing missing data using `na.omit()`. Finally, fix the value of onset that was incorrectly coded as a 4. You may assume for the time being that it was supposed to be a 1.

You may not use canned R functions for this problem except to check answers.

1.A) Provide the appropriate stochastic component, systematic component, and independence assumptions for the logit model used in the paper.

I will use i to denote country-years, but note that it is common to use two subscripts for panel data, i.e. data that consists of a cross-section of units, each observed over multiple time periods. I have not done so here because the panels are unbalanced (some panels have more years than others), so for example, the product in the likelihood would have to be over countries $i \in (1, N)$ and country specific timeframes, $t_i \in (1, T_i)$.

The stochastic component is $Y_i \sim Bernoulli(\pi_i)$

The systematic component is $\pi_i = \frac{1}{(1+e^{-X_i\beta})}$

All observations are independent: $Y_i \perp Y_j$ for $i \neq j$, conditional on X. In particular, this implies that observations are spatially and temporally uncorrelated, conditional on X.

1.B) Derive the log-likelihood for the model parameters, $\ell(\beta|Y_i, X_i, n)$.

$$
\begin{aligned}
L(\pi_i|Y_i, X_i, n) &\propto \prod_i^N (\pi_i)^{y_i}(1-\pi_i)^{(1-y_i)} \\
\ell(\pi_i|Y_i, n) &\propto \sum_i^N y_i ln(\pi_i) + (1-y_i) ln(1-\pi_i) \\
\ell(\beta|Y_i, X_i, n) &\propto \sum_i^N -y_i ln(1+e^{-X_i\beta}) + (1-y_i) ln(1-\frac{1}{1+e^{-X_i\beta}}) \\
&\propto \sum_i^N -y_i ln(1+e^{-X_i\beta}) + (1-y_i) ln(\frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}) \\
&\propto \sum_i^N (1-y_i) ln(e^{-X_i\beta}) - (1-y_i) ln(1+e^{-X_i\beta}) - y_i ln(1+e^{-X_i\beta}) \\
&\propto \sum_i^N (1-y_i) ln(e^{-X_i\beta}) - ln(1+e^{-X_i\beta}) + y_i ln(1+e^{-X_i\beta}) - y_i ln(1+e^{-X_i\beta}) \\
&\propto \sum_i^N (1-y_i) ln(e^{-X_i\beta}) - ln(1+e^{-X_i\beta}) \\
&\propto \sum_i^N ln(e^{-X_i\beta(1-y_i)}) - ln(1+e^{-X_i\beta}) \\
&\propto \sum_i^N -X_i\beta(1-y_i) - ln(1+e^{-X_i\beta}) \\
&\propto -\sum_i^N X_i\beta(1-y_i) + ln(1+e^{-X_i\beta})
\end{aligned}
$$

1.C) Critically evaluate one modelling assumption made by the authors and explain your reasoning.

There are a number of sensible criticisms that one might articulate here, including: the independence assumption for units across time or cross-sectionally is implausible; the measurement of onset is a little strange because years with ongoing full-scale civil wars are consider non-onsets (rather than, say missing data or some third category); and the completely linear function form for the systematic component likely deserves more exploration.

1.D) Write a function in R to implement the log-likelihood you found in 1.B.

Depending on how much you simplified, you might get different answers

```
ll.logit <- function(beta, X, Y) {
 -sum(X%*%beta*(1-Y) + log(1 + exp(-X%*%beta)))

}
```

1.E) Replicate Model 1 from Table 1, using your function from 1.D and `optim()` with `method = "BFGS"`. Be careful in selecting the starting values for your optimization (a vector of zeroes should be adequate). Report selected coefficient estimates and their standard errors.

```
library(foreign)
data <- read.dta("fearondata.dta")
fearon <- data[c("onset","warl","gdpenl","lpopl1","lmtnest","
   ncontig","Oil",
                 "nwstate","instab","polity2l","ethfrac","
                    relfrac")]
fearon <- na.omit(fearon)
fearon["onset"] <- replace(fearon["onset"],fearon["onset"
   ]==4,1)


Y <- as.matrix(fearon[,1])
X <- cbind(1,as.matrix(fearon[,-1]))
opt <- optim(par = rep(0,12),
             Y = Y,
             X = X,
             fn = ll.logit,
             method = "BFGS",
             control = list(fnscale = -1, maxit=100000),
             hessian = TRUE)
opt.coefs <- opt$par
opt.ses <- sqrt(diag(solve(-opt$hessian)))
```

	Coefficient	Std. Error
Intercept	-6.73	0.74
warl	-0.95	0.31
gdpenl	-0.34	0.07
lpopl1	0.26	0.07
lmtnest	0.22	0.08
ncontig	0.44	0.27
Oil	0.86	0.28
nwstate	1.71	0.34
instab	0.62	0.24
polity2l	0.02	0.02
ethfrac	0.17	0.37
relfrac	0.29	0.51

Table 1: Logit coefficients from 1.E

1.F) Using your estimates from 1.E what is the predicted probability of civil war for oil exporters, with all other covariates are held at their median?

The predicted probability is approximately 0.0273.

```
# Set all covariates at median and set oil to 1
setx <- apply(X, 2, median)
setx["Oil"] <- 1
# Calculate predicted probabilities
pred.prob <- 1/(1+exp(-setx%*%opt.coefs))
pred.prob
```

1.G) Conduct a likelihood ratio test which compares the unrestricted model (i.e. the full specification from Model 1) with a restricted model which excludes the ethnic and religious fractionalization variables. Treating the restricted model as your null hypothesis, what is your test statistic and what do you conclude (at the $\alpha = .05$ rejection level)?

```
# Unestricted model
unrestricted  <- optim(par = rep(0,12),
                       Y = Y,
                       X = X,
                       fn = ll.logit,
                       method = "BFGS",
                       control = list(fnscale = -1),
                       hessian = TRUE)

# Restricted Model
X.rest <- cbind(1,as.matrix(fearon[,c(-1,-11,-12)]))
restricted <- optim(par = rep(0,ncol(X.rest)),
                    Y = Y,
                    X = X.rest,
                    fn = ll.logit,
                    method = "BFGS",
                    control = list(fnscale = -1),
                    hessian = TRUE)

r <- 2*(unrestricted$value-restricted$value)
r

#p-value
1 - pchisq(r, df = 2)
```

The p-value of our likelihood ratio test is 0.712. We fail to reject the null model in favor of the unrestricted model at the 0.05 level.

1.H You want to get a rough sense of how well the model fits the data. Using your estimates from 1.E, calculate your predicted probabilities $\hat{\pi}_i$ for each observation. Compute the Brier (1950) score, which is a metric for assessing model fit, defined as

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - Y_i)^2$$

```
# Calculate predicted probabilities
fitted.probs <- 1/(1+exp(-X%*%opt.coefs))

brier.logit <- (1/length(fitted.probs))*sum((fitted.probs - Y)
    ^2)
```

The brier score for the logit model is about 0.015875.

Problem 2: Different Link Functions

In this problem, we are going to investigate what happens to our model when we choose another link function to connect our covariates to the Bernoulli probability. Still assume that we observe n observations with outcome Y_i and covariates X_i. Each $Y_i \sim$ Bernoulli(π_i), but instead of the logit link, we assume that

$$\pi_i = 1 - \exp\left[-\exp(X_i\beta)\right]$$

2.A) Derive the log-likelihood for the model parameters, $\ell(\beta|Y_i, X_i, n)$

$$\begin{aligned}
L(\pi_i|Y_i, X_i, n) &\propto \prod_i^N (\pi_i)^{y_i}(1-\pi_i)^{(1-y_i)} \\
\ell(\pi_i|Y_i, n) &\propto \sum_i^N y_i ln(\pi_i) + (1-y_i)ln(1-\pi_i) \\
\ell(\beta|Y_i, X_i, n) &\propto \sum_i^N y_i ln(1-\exp\left[-\exp(X_i\beta)\right]) + (1-y_i)ln(\exp\left[-\exp(X_i\beta)\right]) \\
&\propto \sum_i^N y_i ln(1-\exp\left[-\exp(X_i\beta)\right]) - (1-y_i)\left[\exp(X_i\beta)\right]
\end{aligned}$$

2.B) Write a function in R to implement the log-likelihood you found in 2.A.

```
ll.cloglog <- function(beta, X, Y) {
  sum(Y * log(1 - exp(-exp(X%*%beta))) - (1 - Y)*exp(X%*%beta)
      )
}
```

2.C) Re-estimate Model 1 from Table 1 of Fearon and Laitin (2003) using your new function from 2.B and `optim()` with `method = "BFGS"`. Note that you will likely need to use different starting values to achieve algorithm convergence – we suggest using 0.15 for all of the β parameters as good starting values. Try different starting values if you obtain an error. Report selected coefficient estimates and their standard errors.

```
## 1.E
library(foreign)
data <- read.dta("fearondata.dta")
fearon <- data[c("onset","warl","gdpenl","lpopl1","lmtnest","
   ncontig","Oil",
                 "nwstate","instab","polity2l","ethfrac","
                    relfrac")]
fearon <- na.omit(fearon)
fearon["onset"] <- replace(fearon["onset"],fearon["onset"
   ]==4,1)


Y <- as.matrix(fearon[,1])
X <- cbind(1,as.matrix(fearon[,-1]))
opt.cloglog <- optim(par = rep(0.15,12),
             Y = Y,
             X = X,
             fn = ll.cloglog,
             method = "BFGS",
             control = list(fnscale = -1, maxit=100000),
             hessian = TRUE)
opt.coefs.cloglog <- opt.cloglog$par
opt.coefs.cloglog
opt.ses.cloglog <- sqrt(diag(solve(-opt.cloglog$hessian)))
```

	Coefficient	Std. Error
Intercept	-6.68	0.72
warl	-0.93	0.31
gdpenl	-0.34	0.07
lpopl1	0.26	0.07
lmtnest	0.21	0.08
ncontig	0.43	0.27
Oil	0.84	0.27
nwstate	1.65	0.32
instab	0.60	0.23
polity2l	0.02	0.02
ethfrac	0.16	0.37
relfrac	0.28	0.50

2.D) Using your estimates from 2.C what is the predicted probability of civil war for oil exporters, with all other covariates are held at their median?

The predicted probability is about 0.272

```
# Set all covariates at median and set oil to 1
setx <- apply(X, 2, median)
setx["Oil"] <- 1
# Calculate predicted probabilities
pred.prob.cloglog <- 1 - exp(-exp(setx%*%opt.coefs.cloglog))
pred.prob.cloglog
```

2.E) You want to compare the fit using your new link function to the fit from your logit model. Using your estimates from 2.C, calculate your predicted probabilities $\hat{\pi}_i$ for each observation. Compute the Brier (1950) score, which is a metric for assessing model fit, defined as

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{\pi}_i - Y_i)^2$$

The Brier score for the new model is 0.015870

```
# Calculate predicted probabilities
fitted.probs.cloglog <- 1 - exp(-exp(X%*%opt.coefs.cloglog))

brier.cloglog <- (1/length(fitted.probs.cloglog))*sum((fitted.
    probs.cloglog - Y)^2)
```

2.F) Based on your answers to 1.G/1.H and 2.D/2.E, do the two models give qualitatively different results? Which model should a researcher choose (if there is a clear winner) and why?

The two models give essentially the same results, despite the differences in model specification. The predicted probabilities are almost exactly the same, and the Brier scores are largely identical (plus/minus some noise), suggesting that the models perform more or less the same in terms of in-sample predictive power. Regardless of which model a researcher chooses, they will get essentially the same substantive results for this particular case.

R Code

Please submit all your code for this assignment as a .R file. Your code should be clean, commented, and executable without error.