Advanced Quantitative Methods in Political Science
Spring 2016
Thomas Gschwend & Tilko Swalve

Homework 4: Binary Outcomes

1 Logit, Probit, and the Linear Probability Model

Ward et al. (2010) reanalyze two canonical statistical models of civil war onset. The first one is published in an often-cited paper by Fearon and Laitin (2003) and the second, similarly important one, by Collier and Hoeffler (2004).

You can find the data sets as used by Ward et al. (2010) on `ILIAS`. The data set for the Fearon and Laitin model is `fl.three.RData`, the data set for the Collier and Hoeffler model is `ch.RData`. Here is an overview of the variables in `fl.three.RData`:

Variable	Dataset
Onset of Civil War	`onset`
Prior War	`warl`
GDP per capita	`gdpenl`
Population	`lpopl1`
Mountainous Terrain	`lmtnest`
Non-contiguous State	`ncontig`
Oil Exporter	`Oil`
New State	`nwstate`
Instability	`instab`
Democracy	`polity21`
Ethnic Fractionalization	`ethfrac`
Religious Fractionalization	`relfrac`

Table 1: Variable included in Fearon and Laitin (2003) model

1. Run a logit, a probit, and a linear probability model (LPM) of the Fearon and Laitin model (see also Ward et al. (2010, p.2)).

2. Except for the LPM, the coefficients are hard to interpret and therefore we are typically interested in the marginal (or partial) effect of increasing a particular variable, holding others constant. Plot the marginal effect of GDP per capita for

the three models holding all other variables at their mean values (no confidence intervals required for now).

3. In a small paragraph, compare the marginal effect plots of the three models with each other.

2 An Alternative Link Function

An alternative to the logit or probit model is the complementary log-log model. The link function of this model is

$$\Pr(y_i = 1) = 1 - \exp(-\exp(X_i\beta)). \tag{1}$$

1. Use the `curve` command to plot the linear, logit, probit, and the complementary log-log link in a single plot and compare the models in a short paragraph.

2. Implement the complementary log-log model in `R`. To do so, you will need to change the link in the logit or probit log-likelihood function.

3. Estimate the Fearon and Laitin model from exercise 1, present the coefficients, and plot the marginal effect of GDP per capita on civil war onset holding all other variables at their respective means. In a small paragraph compare the results to the three models you estimated in exercise 1.

3 Model Selection Issues

A popular tool to select models is the ROC curve. The ROC curve plots the proportion of 1's correctly predicted (true-positive rate) against the proportion of 0's correctly predicted (true-negative rate). Intuitively, a model that only predicts 1's will predict all instances in the data that are actually equal to 1 correctly, but will make many mistakes by doing so. On the other hand, a model that only predicts 0's never mistakenly predicts a 1 where there is non in the data, but it also never identifies a 1.

To make a prediction, the model needs a threshold probability. This specifies a number between 0 and 1 that separates 0 from 1 predictions given predicted probabilities. Here is some `R`-code that calculates the true-positive and the false-positive rate for a threshold of 0.3 with an example data set.

```r
load('HW4Example.Rdata')

m1 <- glm(y ~ X1 + X2 + X3,data=HW4,family=binomial())
covar <- cbind(1,HW4$X1,HW4$X2,HW4$X3)

# Predicted Probability
mu <- covar %*% m1$coef
p <- 1/(1 + exp(-mu))

# Threshold
threshold <- 0.3

# Prediction
pred <- NULL
pred[p >= threshold] <- 1
pred[p < threshold] <- 0

# True Positive Rate
tpr <- sum(HW4$y==1 & pred==1)/sum(HW4$y==1)
# True Negative Rate
tnr <- sum(HW4$y==0 & pred==0)/sum(HW4$y==0)
```

1. Write a function that calculates the true-positive and the true-negative rate for a range of thresholds from 0 to 1 in steps of 0.05.

2. Use this function to compare the Laitin and Fearon model with the Collier and Hoeffler model (see variables below). Which model do you prefer?

4 Specification Issues

We are well familiar with the omitted variable bias in the linear model. Omitted variable bias was present if an unobserved regressor was correlated with both the dependent and an independent variable. In that case, we know that our estimates will be inconsistent. The same issue arises in logit and probit models.

However, another problem can occur in logit and probit models known as *neglected heterogeneity*. Estimates can be biased even if the unobserved (omitted) variable and the

Variable	Dataset
Onset of Civil War	`warsa`
Commodity Dependence	`sxp`
Squared Commodity Dependence	`sxp2`
Male Secondary Schooling	`secm`
GDP Growth	`gy1`
Peace Duration	`peace`
Geographic Dispersion	`geogia`
Population	`lnpop`
Social Fractionalization	`frac4590`
Ethnic Dominance	`etdo4590`

Table 2: Variable included in Collier and Hoeffler (2004) model

observed variable are uncorrelated! In this exercise you will explore the nature of this bias using Monte Carlo simulation.

Use the following piece of R code to generate some fake data:

```
set.seed(123)
n <- 10000
X1 <- rnorm(n,0,1)
X2 <- rnorm(n,0,3)
beta0 <- 0.5
beta1 <- 1
beta2 <- 1.5

mu <- beta0 + beta1*X1 + beta2*X2
p <- 1/(1 + exp(-mu))

y <- rbinom(n,1,p)
```

1. Check that X_1 and X_2 are not correlated.
2. Using the fake data, specify two logit models: a full model including $X1$ and $X2$ and a restricted model only including X_1.
3. Specify a full and a restricted model as linear probability models.
4. Conduct Monte Carlo simulations comparing the full and the restricted model for the logit and the linear probability model.
5. Describe your results in a short paragraph and two plots.

6. Choose different values for the coefficient and the variance of the unobserved regressor. How are your results influenced by these different values?

5 Final Paper Proposal

Please provide a short (< 150 words) description of what you are going to do (data and methods) in your final paper.

References

Collier, P. and Hoeffler, A. (2004). Greed and grievance in civil war. *Oxford Economic Papers*, 56(4):563–595.

Fearon, J. D. and Laitin, D. D. (2003). Ethnicity , Insurgency , and Civil War. *American Political Science Review*, 97(1):75–90.

Ward, M. D., Greenhill, B. D., and Bakke, K. M. (2010). The perils of policy by p-value: Predicting civil conflicts. *Journal of Peace Research*, 47(4):363–375.