

Advanced Quantitative Methods

Homework 3: Causal Inference & Maximum Likelihood

Johan A. Elkind
jos.elkind@ucd.ie

Due 5 April 2017, 5pm

Please submit by email in PDF format. Add R code in a separate .R file. Note that if you do not use Latex, there is an “Equation” entry in Microsoft Word under “Insert” that will allow you to include matrices and other mathematical equations.

Percentages with an asterisk indicate that positive rather than negative marking will be applied.

(5*%) of the grade is used for an overall evaluation of the clarity and presentation of your code.

Data

Just as the previous, this homework is based on the replication data for Ross and Voeten (2016), which you can access at their Harvard Dataverse record.¹

The screenshot shows a list of data files on the Harvard Dataverse platform. Each entry includes the file name, format, size, date, and number of downloads. For the 'ReplicationdataRossVoeten.tab' entry, the 'Download' button is clicked, revealing a dropdown menu with several options. The 'Original File Format (Stata 13 Binary)' option is highlighted with a red circle.

The “explore” button might also be very helpful to get a feel for the data. It is of course advisable to have a quick glance at their paper as well.

¹<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BON24T>

Questions

1. Based on Ross and Voeten (2016), we will be investigating the relationship between democracy and membership of “structured” international organisations using the following model:

$$\text{strucint}_i = \beta_1 + \beta_2 \text{democracy}_i + \beta_3 \text{oilexp}_i + \beta_4 \text{lngdp}_i + \beta_5 \text{lnpop}_i + \varepsilon_i,$$

where our key independent variable, the treatment, is **democracy**.

- (a) Select only cases for the year 2001.
- (b) Perform the above regression.
- (c) Repeat, adding **Intradegdp**—the log of trade over GDP—as an additional control variable.
- (d) (15%) Provide a motivation (approximately 200 words) why this (**Intradegdp**) is a suitable control variable.
- (e) Instead of adding the control variables to the regression, we can also use a matching algorithm. Note that we use the `model.frame()` function to handle missing data. The `MatchIt` library in R provides a function for propensity score matching. (5%) Execute the following code:

```
library(MatchIt)

f <- strucint ~ democracy + oilexp + lngdp + lnpop
mf <- model.frame(f, data)

msample <- matchit(democracy ~ oilexp + lngdp + lnpop, mf,
  method = "nearest", replace = TRUE)
mm <- lm(strucint ~ democracy, data = match.data(msample))
```

- (f) (5%) Repeat, but add the control variables to the regression model as well.
 - (g) (10%) Present a regression table with all four regressions.
 - (h) (10%) Interpret the results (approximately 300 words).
2. We continue with the same regression model (in Question 2b).
 - (a) The dependent variable in the above model is a count variable, which has by definition only positive values. The normal distribution is therefore not strictly appropriate. The poisson distribution is a distribution which only takes positive, integer values (0, 1, 2, ...), and has the following density (mass) function:

$$f(y|\lambda) = \frac{e^{-\lambda} \lambda^y}{y!},$$

where the parameter λ reflects both the variance and the expected value.² This leads to the following log-likelihood function:

$$\ell(\lambda|y) = \sum_{i=1}^N (y_i \log \lambda - \log(y_i!)) - N\lambda.$$

Since λ is the expected value and since this has to be positive by definition, we typically model the link function as $\lambda = e^{\mathbf{X}\beta}$.³ (10%) Finish the following implementation of the log-likelihood function in R:

```
loglik <- function(beta, X, y) {  
  
  lambda <- exp(X %*% beta)  
  
  # here goes the loglikelihood function  
  
}
```

- (b) (10%) Using the `optim()` function, estimate the model parameters for the above regression model. Note that our implementation cannot handle missing data, so we need to generate \mathbf{X} and \mathbf{y} explicitly first:

```
X <- model.matrix(f, data)  
y <- model.response(mf)
```

In addition to the parameters as in the slides, you should also pass `method = "BFGS"` to the `optim()` function.

- (c) Extract the estimated coefficients and (10%) their associated standard errors.
(d) (10%) Perform a z -test for each of the coefficients.⁴
(e) (10%) Briefly interpret the results (approximately 200 words).

²Note that in R, $y_i!$ is `factorial(y)`.

³Note that the Poisson distribution is relatively inflexible, because the same parameter regulates both the mean and the variance. Alternative distributions such as the negative binomial are available, but not part of this exercise.

⁴Note that all results in this question can be verified using R's built-in function: `summary(glm(f, data, poisson("log")))`.

Grade conversion scheme

Score	Grade		Score	Grade		Score	Grade		Score	Grade	
	UCD	TCD		UCD	TCD		UCD	TCD		UCD	TCD
97-100%	A+	A+	85-87%	B	B	74-76%	C-	C	54-64%	E+	D
94-96%	A	A	83-84%	B-	B	71-73%	D+	C	44-53%	E	D
91-93%	A-	A	80-82%	C+	C+	68-70%	D	C	33-43%	E-	D
88-90%	B+	B+	77-79%	C	C	65-67%	D-	C	0-32%	F	F

Ross, Michael L. and Erik Voeten. 2016. "Oil and international cooperation." *International Studies Quarterly* 60(1):85–97.

URL: <https://www.sscnet.ucla.edu/polisci/faculty/ross/publications.html>