Advanced Quantitative Methods
Homework 3: causal inference & maximum likelihood

Johan A. Elkink
jos.elkink@ucd.ie

Due April 30, 2018, 5pm

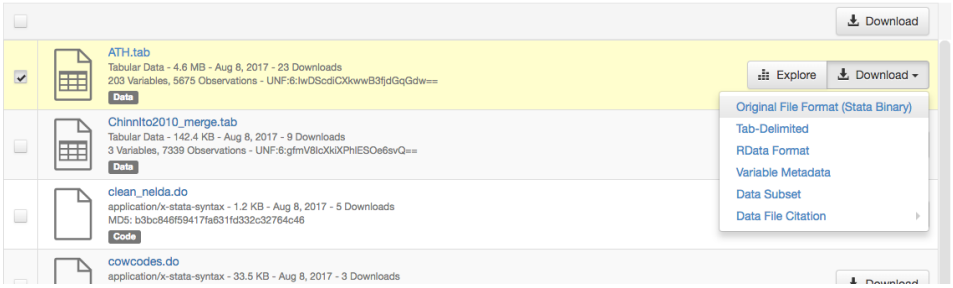
Please submit by email in PDF format. Add R code in a separate `.R` *file, or SPSS code in a separate* `.sps` *file, or Stata code in a separate* `.do` *file, or the code for any other package you use separately. Alternatively, you can integrate R code and responses in R Markdown and submit both in PDF format (easiest is to use HTML format and then in the web browser save as PDF).*

Please note the revised submission date.

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

Data

This homework is based on the replication data for **?**, which you can access at their Harvard Dataverse record—probably the first study you find when searching for the keyword "remittance".[1]



The "explore" button might also be very helpful to get a feel for the data.

Questions

1. We will be investigating the relationship between democracy in neighbouring countries and democracy within a country itself, using the Polity IV democracy score, which runs from -10 for autocracies to +10 for democracies.

[1] `https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/NFMGFD`

(a) Create a variable **neighbor_democracy** that is one if the average Polity IV score among neighbours (**neighbor_pol4_polity2**) is greater than 0 and zero otherwise.

(b) Create a variable **oilproducing** that is one if **ross_oilproduction** greater than 0 and zero otherwise.

(c) We will use a trick of self-merging to generate lagged versions of all variables (in this code I assume the data set is called "ath"):

```
ath$lastyear <- ath$year - 1
ath <- merge(ath, ath, by.x = c("cowcode", "lastyear"),
        by.y = c("cowcode", "year"), suffixes = c("", ".lag"),
        all.x = TRUE)
```

(d) (5%) Run a baseline regression, regressing the Polity IV score (**pol4_polity2**) on the neighbours dummy variable (**neighbor_democracy.lag**), including a lagged dependent variable (**pol4_polity2.lag**) in the model.

(e) (5%) Using the `lmer()` function, estimate the model adding country (**cowcode**) random effects (i.e. random intercepts).

(f) (5%) Add control variables for oil producing countries (**oilproducing.lag**), log of GDP per capita (**mad_lgdppc.lag**), and log of population size (**mad_lpop.lag**).[2]

2. (10%) Provide a motivation (approximately 200 words) why (**mad_lgdppc.lag**) is a suitable control variable.

3. We repeat the above analysis, but using matching instead of the overall sample.

   (a) (10%) Instead of adding the control variables to the regression, use a nearest neighbour matching algorithm to get a matched sample. Note that the lagged dependent variable is not a control variable and should therefore not be used in the matching. You can use the `model.frame()` function to extract a data set with all missing values removed from the above regression output.

   (b) (10%) Repeat the three regressions on the matched sample.

   (c) (5%) Present a regression table with all six regressions.

   (d) (15%) Interpret the results (approximately 400 words). Aside from the substantive interpretation, discuss in particular how the estimate for the effect of neighbouring democracies changes between the different models.

4. When looking at a plot of the residuals, you will notice that these are far from normally distributed. This might be in part due to the truncated nature of the democracy scale. We will therefore assume a truncated normal distribution instead of a normal distribution for the errors and re-estimate the linear model (ignoring the random effects specification).

[2]Note that the "l" in the variable name indicates that these are logged values as well, so there is no need to take the logarithm explicitly—this is already done.

The loglikelihood function for a truncated normal distribution is as follows:

$$-\frac{1}{2}\left(n\log(2\pi)+n\log(\sigma^2)+\frac{1}{\sigma^2}\sum_i^n e_i^2\right)-n\log(\sigma)-\sum_i^n\left[\log(\Phi_b(X\beta,\sigma^2)-\Phi_a(X\beta,\sigma^2)\right],$$

where $e = y - X\beta$ and $\Phi_b(X\beta, \sigma^2)$ is the cumulative normal distribution with mean $X\beta$ and variance σ^2, evaluated at upper limit b, which (assuming the upper bound of the truncation b is called `upp`) translates into R as:

```
pnorm(upp, X %*% beta, sqrt(s2))
```

(a) (10%) Complete the following loglikelihood function:

```
loglik <- function(theta, y, X, low, upp) {


}
```

(b) We define out variables as follows, assuming the last model estimated (the output from `lmer()`) is called `md16`:

```
mf <- model.frame(md16)
y <- mf$pol4_polity2
X <- as.matrix(cbind(1, mf[, 2:6]))
```

(5%) Test the loglikelihood function when all θ values are zero. (I obtain -38921.53.)

(c) (5%) Estimate the model using the `optim()` function and report the obtained coefficient estimates.

(d) (10%) Calculate standard errors for all parameters.

(e) (5%) Calculate z-scores and associated p-values for all parameters.

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