Proposal: The Effect of U.S. Military Aid on FDI

Ben Horvath

November 1, 2018

Load libraries:

```r
library(dplyr)
library(ggplot2)
library(Hmisc)
library(knitr)
library(readxl)
library(stringr)
library(tidyr)
```

Introduction

I took a 200-level 'methods in political science' class in the spring semester of 2009, as part of an undergraduate degree in political science. One assignment was to write a proposal for empirical research, though we weren't required to necessarily carry out the analysis.

Biglaiser and DeRouen (2007) and Little and Leblang (2004) had found that 'the presence of U.S. troops serves as a "catalyst" for U.S. outgoing foreign direct investment (FDI), that is, FDI follows the flag.' My proposal, 'The Effect of U.S. Military Aid on Foreign Direct Investment Decisions,' was to test if U.S. arms exports had a similar effect on FDI as U.S. troops. I located the appropriate data, and noted a number of other variables that would need to be controlled for: alliances, the presence of conflict, the Cold War and outliers like Vietnam, regime type (democratic, autocratic, etc.), and population size.

For methodology, I wrote, simply, 'To test these associations, some kind of regression would be used, with an appropriate significance test.' I don't believe I really knew what a regression was, I'd just noticed how popular it was in the political science literature.

Nine years later, I am much more sophisticated statistically, and would like to see how my college sophomore intuition faired.

References

- Glenn Biglaiser and Karl DeRouen, Jr. (2007), 'Following the flag: Troop deployment and US foreign direct investment,' *International Studies Quarterly* 51, no. 4: 835-854.
- Andrea Little and David Leblang (2004), 'Military securities: Financial flows and the deployment of US troops,' in *Annual Meeting of the American Political Science Association*, pp. 2-5.

Research Questions

What is the effect of an increase in U.S. arms exports to a country's incoming U.S. foreign direct investment?

Cases

Each row will be attributes associated with a single year for a single country: `(year, country)`.

Note: I realize the basic regression we're going to perform is not ideal for this kind of cross-sectional longitudinal data set. It clearly violates, at least, the assumption of independence between observations. However, I'd like to see why it doesn't work for myself, on a concrete dataset I understand. I'd also like to compare this basic regression's performance against the 'proper' way as well as the methodology of the studies referenced above, as a bonus.

Response and Explanatory Variables

The response variable is incoming U.S. FDI from a country, measured in USD.

The explanatory variable we are most interested in is U.S. military aid (probably lagged a year).

The studies referenced above include a few control variables, including population, existence of a conflict in that year and country, type of regime (democracy, dictatorship, etc.), alliance statuses, distance between countries, and GDP. These variables will also be lagged a year for modeling.

All of this data is easily assembled, if you know where to look, so I would like to include those as well.

Data Sources

- **FDI**. The OECD provides FDI data for U.S. outflows on its website, from 2003 to 2013. These years will have to bound this study temporally: https://stats.oecd.org/index.aspx?DataSetCode=FDI_FLOW_PARTNER
- **Arms transfers**. The Stockholm International Peace Research Institute maintains a database of arms transfers: https://www.sipri.org/databases/armstransfers. The value has been 'normalized' by the researchers themselves to account for fluctuations in the market value of weapons as well as allowing comparability between, e.g., 100 assault rifles and 2 large artilleries.
- **Yearly Population**. This is available for most countries on a yearly basis via the UN: https://population.un.org/wpp/Download/Standard/Population/
- **Presence of Conflict**. Political scientists testing hypotheses on armed conflict frequently make use of the Armed Conflict dataset, available at: http://ucdp.uu.se/downloads/#d3. This dataset contains a lot of data, but I am just going to use a dichotomous variable: 0 for no conflict, 1 for conflict.
- **Regime Type**. This is available in one of the most popular political science datasets, the Polity dataset. It encodes regime type in a range from perfectly democratic to perfectly autocratic for most countries from 1800 on. Specifically I will use the Polity2 variable: http://www.systemicpeace.org/inscrdata.html. See also the user manual: http://www.systemicpeace.org/inscr/p4manualv2017.pdf.
- **Alliances**. The Correlates of War project maintains another popular data set encoding international alliances in 'dyadic' form a year-to-year basis: http://www.correlatesofwar.org/data-sets/formal-alliances.
- **Distance**. Kristian Gleditsch developed a data set containing the distance between capital cities, which we'll use to proxy distance: http://ksgleditsch.com/data-5.html, using this system of country codes: http://ksgleditsch.com/statelist.html
- **GDP**. Where else but the World Bank?: https://data.worldbank.org/indicator/NY.GDP.MKTP.CD

Data Collection and Preparation

Collecting Data

The goal will be to combine this data in a clean format, for as many countries as possible between 2003 and 2013. Filter the dataset to only include outflow numbers.

FDI

```
fdi <- read.csv('../data/raw/FDI_FLOW_PARTNER_28102018214703504.csv',
                stringsAsFactors=FALSE)
colnames(fdi) <- tolower(colnames(fdi))

# we are only interested in outflow, from the U.S. to other countries
fdi <- fdi %>%
    filter(reporting.country == 'United States',
           type.of.fdi == 'Outward',
           cur == 'USD')
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.4
```

The data also includes various aggregated rows, including regions like GULF ARABIAN COUNTRIES, these must be filtered out as well.

```
`%not in%` <- function (x, table) is.na(match(x, table, nomatch=NA_integer_))

aggregations <- c('ACP countries', 'AFRICA', 'African ACP countries', 'AMERICA', 'ASEAN

fdi <- fdi %>% filter(partner.country %not in% aggregations)
```

Clean up the columns a bit:

```
us_fdi <- fdi %>%
    select(year, partner.country, value) %>%
    mutate(value = value * 1000000) %>%
    arrange(year, partner.country)
colnames(us_fdi) <- c('year', 'country', 'fdi')
rm(fdi)

write.table(us_fdi, '../data/clean/fdi.tsv', row.names=FALSE, sep='\t')
```

```
head(us_fdi)
```

```
##   year             country       fdi
## 1 2003             Albania  -1000000
## 2 2003             Algeria 636000000
## 3 2003             Andorra         0
## 4 2003              Angola -36000000
## 5 2003            Anguilla  -2000000
## 6 2003 Antigua and Barbuda  -4000000
```

Arms Transfers

```
# not really an Excel file!
arms_raw <- read.csv('../data/raw/TIV-Export-USA-2003-2013.csv.xls', skip=10,
                  header=TRUE)
arms_raw$Total <- NULL
colnames(arms_raw)[1] <- 'country'

arms_exports <- arms_raw %>%
    gather(year, arms_exports, X2003:X2013, na.rm=TRUE) %>%
    mutate(year=as.integer(str_remove(year, 'X'))) %>%
    arrange(country, year)
colnames(arms_exports)[3] <- 'arms_exports'
rm(arms_raw)

write.table(arms_exports, '../data/clean/arms_exports.tsv', row.names=FALSE, sep='\t')

head(arms_exports)
```

```
##       country year arms_exports
## 1 Afghanistan 2005           19
## 2 Afghanistan 2007           22
## 3 Afghanistan 2008           78
## 4 Afghanistan 2009          280
## 5 Afghanistan 2010          245
## 6 Afghanistan 2011          520
```

Population

```
pop_raw <- read_xlsx('../data/raw/WPP2017_POP_F01_1_TOTAL_POPULATION_BOTH_SEXES.xlsx',
                     skip=16, col_names=TRUE) %>%
    select(3, `2003`:`2013`)
colnames(pop_raw)[1] <- 'country'

pop_aggs <- c('WORLD', 'More developed regions', 'Less developed regions', 'Least develo

pop <- pop_raw %>%
    filter(country %not in% pop_aggs) %>%
    gather(year, population, `2003`:`2013`) %>%
    mutate(population = population * 1000,
           year = as.numeric(year))

rm(pop_raw)
write.table(pop, '../data/clean/population.tsv', row.names=FALSE, sep='\t')

head(pop)
```

```
## # A tibble: 6 x 3
##   country   year population
##   <chr>    <dbl>      <dbl>
## 1 Burundi   2003    6953113
## 2 Comoros   2003     583211
## 3 Djibouti  2003     758615
## 4 Eritrea   2003    3738265
## 5 Ethiopia  2003   72545144
## 6 Kenya     2003   34130852
```

Conflict

```
conflict <- read.csv('../data/raw/ucdp-prio-acd-181.csv',
                         stringsAsFactors=FALSE) %>%
    select(conflict_id, location, year) %>%
    mutate(conflict = 1) %>%
    arrange(conflict_id, location, year)
colnames(conflict)[2] <- 'country'
```

```
write.table(conflict, '../data/clean/conflict.tsv', row.names=FALSE, sep='\t')

head(conflict)
```

```
##   conflict_id              country year conflict
## 1         200              Bolivia 1946        1
## 2         200              Bolivia 1952        1
## 3         200              Bolivia 1967        1
## 4         201 Cambodia (Kampuchea) 1946        1
## 5         201 Cambodia (Kampuchea) 1947        1
## 6         201 Cambodia (Kampuchea) 1948        1
```

Regime Type

```
regime <- read_xls('../data/raw/p4v2017.xls') %>%
    select(country, year, polity2) %>%
    arrange(country, year)

write.table(regime, '../data/clean/regime.tsv', row.names=FALSE, sep='\t')

head(regime)
```

```
## # A tibble: 6 x 3
##   country      year polity2
##   <chr>       <dbl>   <dbl>
## 1 Afghanistan  1800      -6
## 2 Afghanistan  1801      -6
## 3 Afghanistan  1802      -6
## 4 Afghanistan  1803      -6
## 5 Afghanistan  1804      -6
## 6 Afghanistan  1805      -6
```

Alliances

```
alliances <- read.csv('../data/raw/alliance_v4.1_by_dyad_yearly.csv',
                      stringsAsFactors=FALSE) %>%
    filter(state_name1 == 'United States of America',
           year >= 2003,
           year <= 2013) %>%
    select(state_name2, year) %>%
    mutate(alliance = 1)
colnames(alliances)[1] <- 'country'

write.table(alliances, '../data/clean/alliances.tsv', row.names=FALSE, sep='\t')

head(alliances)
```

```
##   country year alliance
## 1  Canada 2003        1
## 2  Canada 2004        1
## 3  Canada 2005        1
## 4  Canada 2006        1
## 5  Canada 2007        1
## 6  Canada 2008        1
```

Distance

```
countries <- read.table('../data/raw/iisystem.dat', sep='\t',
                        stringsAsFactors=FALSE)

distance <- read.csv('../data/raw/capdist.csv', stringsAsFactors=FALSE) %>%
    filter(ida == 'USA') %>%
    inner_join(countries, by=c('idb'='V2')) %>%
    select(V3, kmdist)
colnames(distance) <- c('country', 'km_dist')

rm(countries)
write.table(distance, '../data/clean/distance.tsv', row.names=FALSE, sep='\t')

head(distance)
```

```
##              country km_dist
## 1             Canada     731
## 2            Bahamas    1623
## 3               Cuba    1813
## 4              Haiti    2286
## 5              Haiti    2286
## 6 Dominican Republic    2358
```

GDP

Extracting two variables, absolute GDP and yearly percentage growth:

```
gdp <- read.csv('../data/raw/API_NY.GDP.MKTP.CD_DS2_en_csv_v2_10203569.csv',
                stringsAsFactors=FALSE, skip=4) %>%
    select(Country.Name, X2002:X2013) %>%
    gather(year, gdp, X2002:X2013) %>%
    mutate(year = as.numeric(str_remove(year, 'X'))) %>%
    arrange(Country.Name, year)
colnames(gdp) <- c('country', 'year', 'gdp')

gdp <- gdp %>%
    group_by(country) %>%
    mutate(gdp_l1 = lag(gdp, n=1, default=NA)) %>%
    mutate(gdp_perc_growth = (gdp - gdp_l1) / gdp_l1) %>%
    filter(year >= 2003, year <= 2013) %>%
    select(country, year, gdp, gdp_perc_growth)

write.table(gdp, '../data/clean/gdp.tsv', sep='\t', row.names=FALSE)

head(gdp)
```

```
## # A tibble: 6 x 4
## # Groups:   country [1]
##   country      year         gdp gdp_perc_growth
##   <chr>       <dbl>       <dbl>           <dbl>
## 1 Afghanistan  2003  4583644246.          0.110
## 2 Afghanistan  2004  5285465686.          0.153
## 3 Afghanistan  2005  6275073572.          0.187
## 4 Afghanistan  2006  7057598407.          0.125
## 5 Afghanistan  2007  9843842455.          0.395
## 6 Afghanistan  2008 10190529882.          0.0352
```

Putting the Data Together

```
df <- us_fdi %>%
    inner_join(pop, by=c('country', 'year')) %>%
    left_join(arms_exports, by=c('country', 'year')) %>%
    left_join(conflict,  by=c('country', 'year')) %>%
    left_join(regime, by=c('country', 'year')) %>%
    left_join(alliances, by=c('country', 'year')) %>%
    left_join(distance, by=c('country')) %>%
    left_join(gdp, by=c('country', 'year')) %>%
    select(-conflict_id) %>%
    arrange(country, year)
```

```
## Warning: Column `country` joining character vector and factor, coercing
## into character vector
```

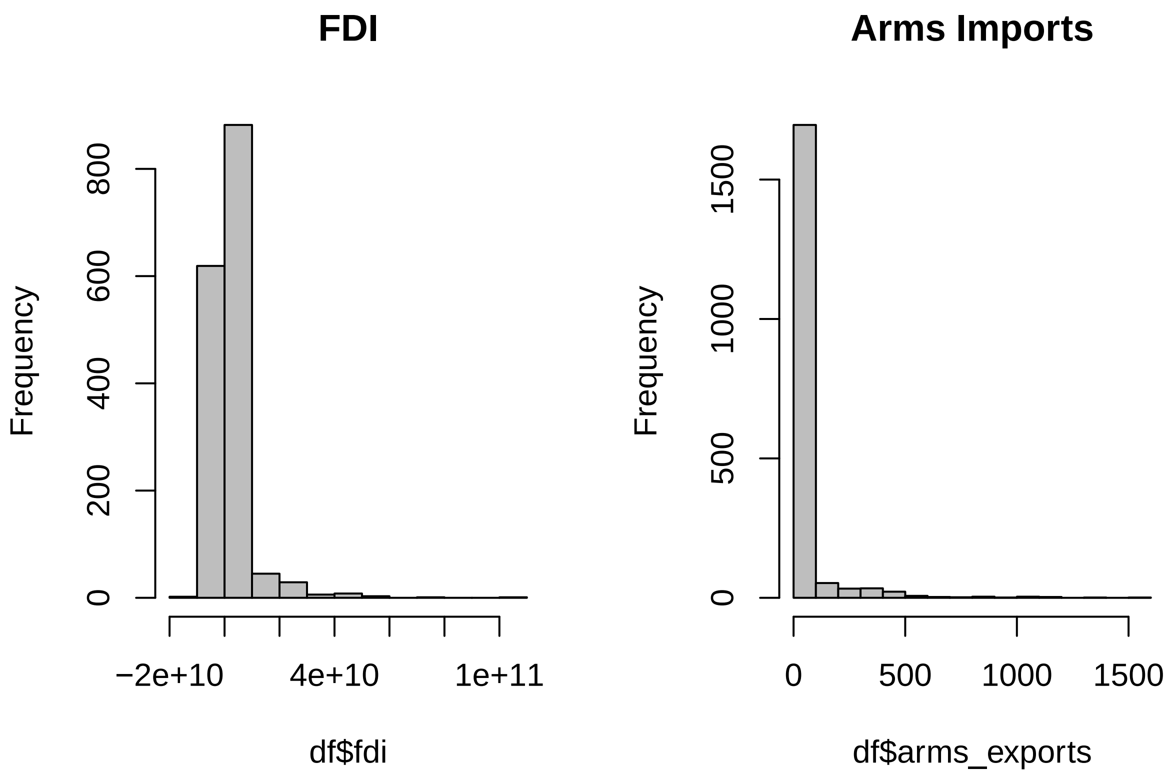
```
# Fill in missing variables to indicate absense of exports, etc.
df <- df %>%
    mutate(arms_exports = replace_na(arms_exports, 0),
           conflict = replace_na(conflict, 0),
           alliance = replace_na(alliance, 0))


write.table(df, '../data/clean/master_dataset.tsv', row.names=FALSE, sep='\t')

head(df)
```

```
##   year     country    fdi population arms_exports conflict polity2
## 1 2004 Afghanistan  0e+00   24118979            0        1      NA
## 2 2005 Afghanistan  0e+00   25070798           19        1      NA
## 3 2006 Afghanistan  0e+00   25893450            0        1      NA
## 4 2007 Afghanistan  0e+00   26616792           22        1      NA
## 5 2008 Afghanistan  0e+00   27294031           78        1      NA
## 6 2009 Afghanistan -1e+06   28004331          280        1      NA
##   alliance km_dist         gdp gdp_perc_growth
## 1        0      NA  5285465686      0.15311429
## 2        0      NA  6275073572      0.18723192
## 3        0      NA  7057598407      0.12470369
## 4        0      NA  9843842455      0.39478643
## 5        0      NA 10190529882      0.03521871
## 6        0      NA 12486943506      0.22534781
```

Summary Statistics

These are histograms of the two main variables of interest, incoming FDI and arms imports from the U.S. Both are about the same shape, with many countries that have little or no FDI or arms imports, and a long right tail.

```
par(mfrow=c(1,2))
hist(df$fdi, main='FDI', col='gray')
hist(df$arms_exports, main='Arms Imports', col='gray')
```

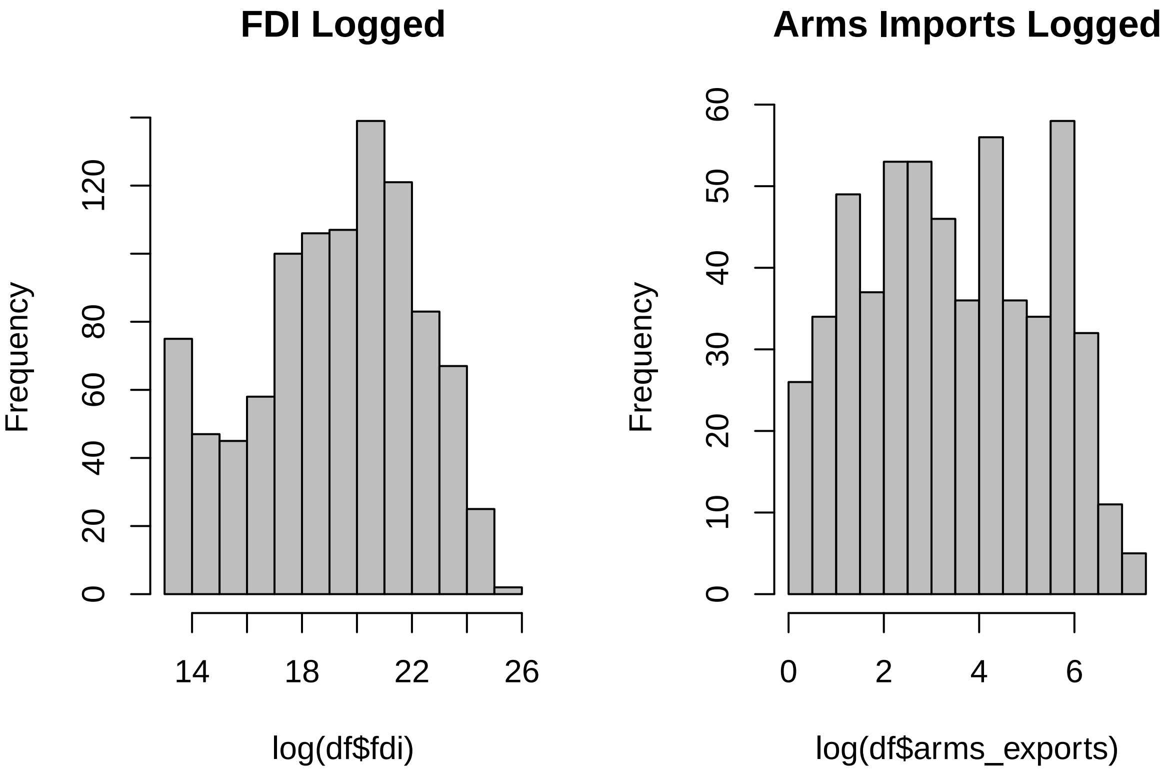


It might make sense for modeling purposes to take the logs of these variables, or at least of FDI—though some values are negative:

```
par(mfrow=c(1,2))
hist(log(df$fdi), main='FDI Logged', col='gray')
```

```
## Warning in log(df$fdi): NaNs produced
```

```
hist(log(df$arms_exports), main='Arms Imports Logged', col='gray')
```



The tremendous skew in FDI is obvious by comparing the mean and the median, almost 2 billion and 10 million, respectively. Standard deviation is also very high, almost 700 million.

```
describe(df$fdi)
```

```
## df$fdi
##          n    missing   distinct       Info       Mean        Gmd
##       1596        268        711      0.994  1.887e+09  3.764e+09
##        .05        .10        .25        .50        .75        .90
## -2.030e+08 -3.700e+07  0.000e+00  1.000e+07  6.820e+08  4.116e+09
##        .95
##  1.236e+10
##
## lowest : -19284000000 -15041000000  -9708000000  -8797000000  -8545000000
## highest:  50184000000  50230000000  51588000000  75007000000 109097000000
```

It occurrs to me it might make sense to standardize FDI by dividing it by a country's population or GDP—something to experiment with.

Just for fun, let's look at the top 10 recipients of FDI from the U.S. (in millions of USD):

```
df %>% group_by(country) %>%
    summarise(fdi = sum(fdi)) %>%
    mutate(fdi = fdi / 1000000) %>%
    arrange(desc(fdi)) %>%
    select(country, fdi) %>%
    top_n(10) %>%
    kable
```

```
## Selecting by fdi
```

country	fdi
Canada	578682
Netherlands	438331
United Kingdom	299169
Luxembourg	224771
Ireland	166329
Bermuda	134972
Switzerland	102596
India	97402
Cayman Islands	92863
Mexico	73762

Most are highly developed nations with large GDPs, half of them in Europe.

Arms exports also has a much larger mean than median, the latter at 0, i.e., over half of these year-country units received no arms from the United States.

```
describe(df$arms_exports)
```

```
## df$arms_exports
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1864        0      190    0.662    37.18    67.87      0.0      0.0
##      .25      .50      .75      .90      .95
##      0.0      0.0      4.0     87.7    272.0
##
## lowest :    0    1    2    3    4, highest: 1027 1110 1114 1389 1526
```

The top 10 recipients of U.S. arms over this time period:

```
df %>% group_by(country) %>%
    summarise(arms_exports = sum(arms_exports )) %>%
    arrange(desc(arms_exports )) %>%
    select(country,arms_exports ) %>%
    top_n(10) %>%
    kable
```

```
## Selecting by arms_exports
```

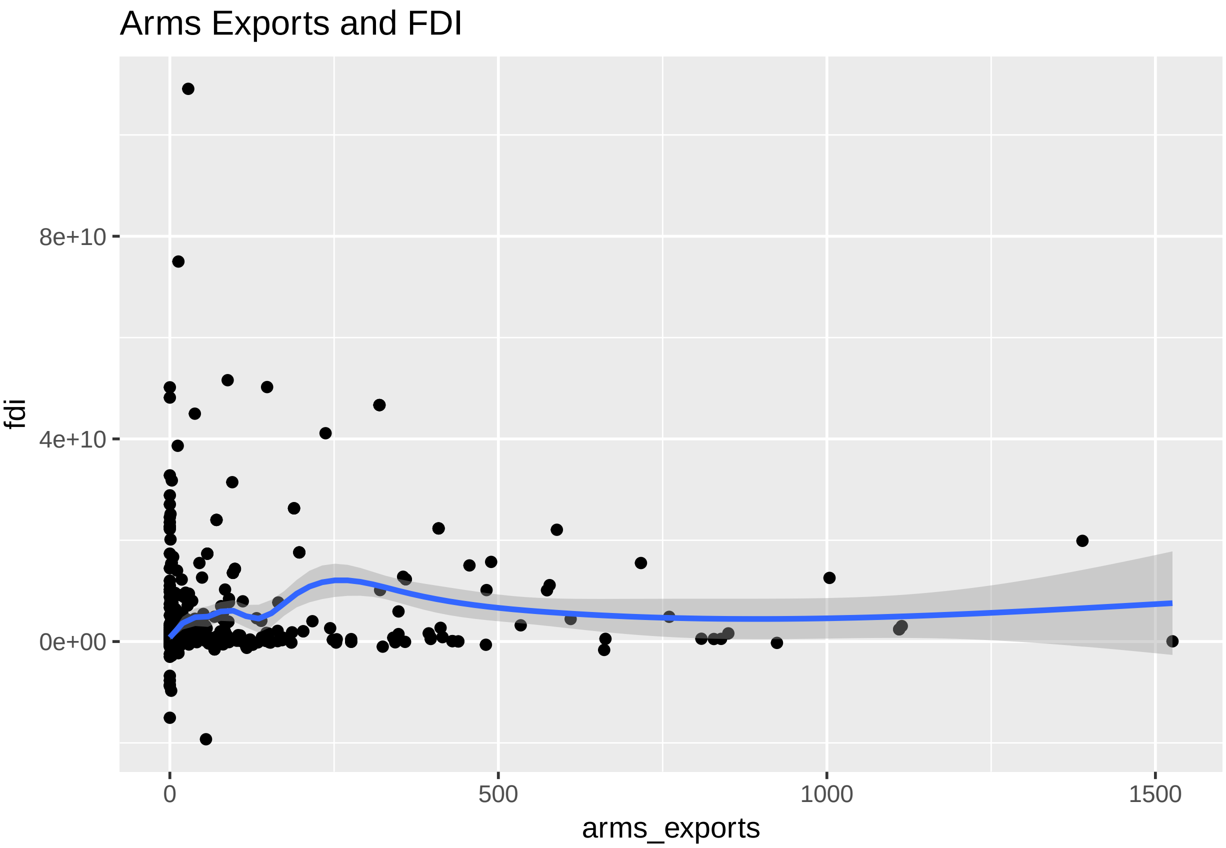
country	arms_exports
Egypt	6434
Israel	6090
Canada	5592
Australia	5214
Turkey	4716
Pakistan	4183
Japan	4099
Greece	3839
Singapore	3439
United Kingdom	2627

This all looks correct. Israel is the largest recipient, and Egypt receives tons of aid under the Camp David Treaty that President Carter negotiated.

Let's look at a scatterplot to view the direct relationship between FDI and arms exports:

```
ggplot(na.omit(df), aes(x=arms_exports, y=fdi)) +
    geom_point() +
    geom_smooth() +
    ggtitle('Arms Exports and FDI')
```

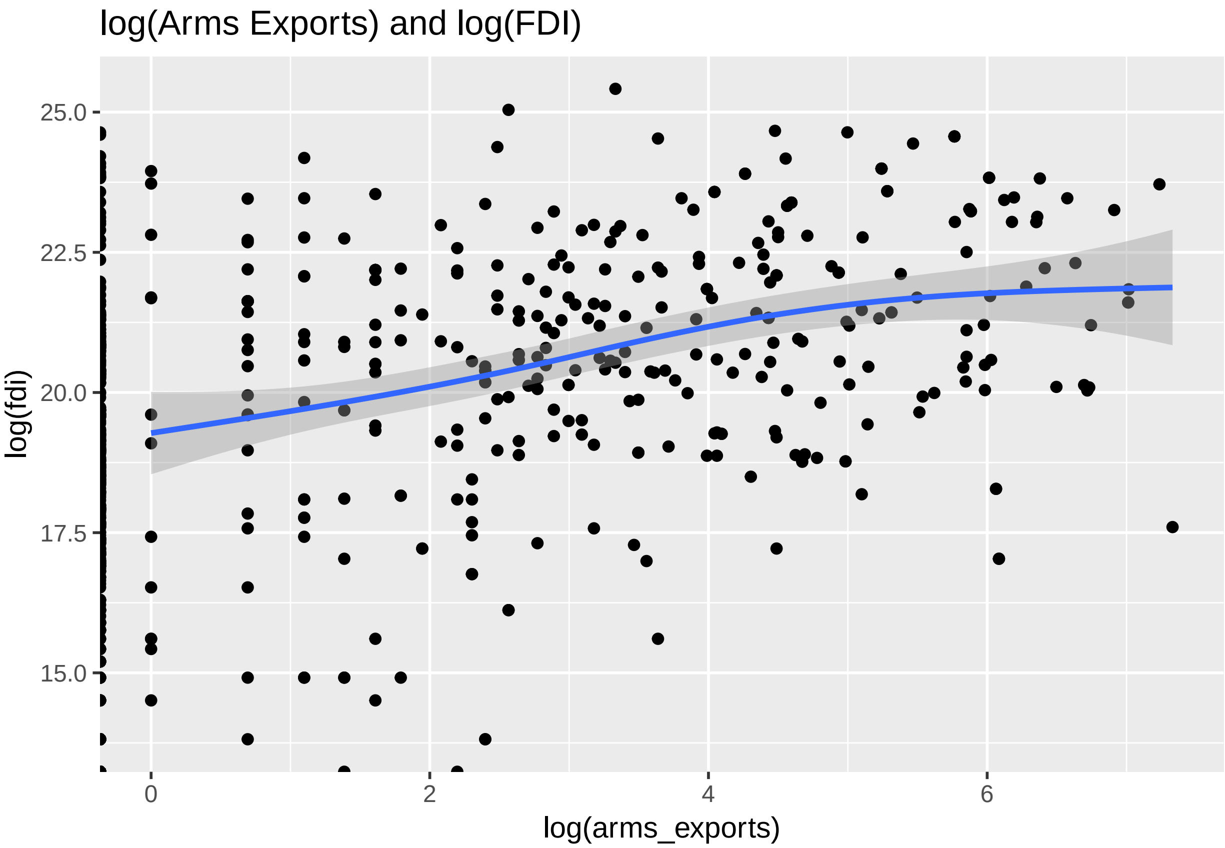
```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



Let's try taking the log of both variables:

```
ggplot(na.omit(df), aes(x=log(arms_exports), y=log(fdi))) +
    geom_point() +
    geom_smooth() +
    ggtitle('log(Arms Exports) and log(FDI)')
```

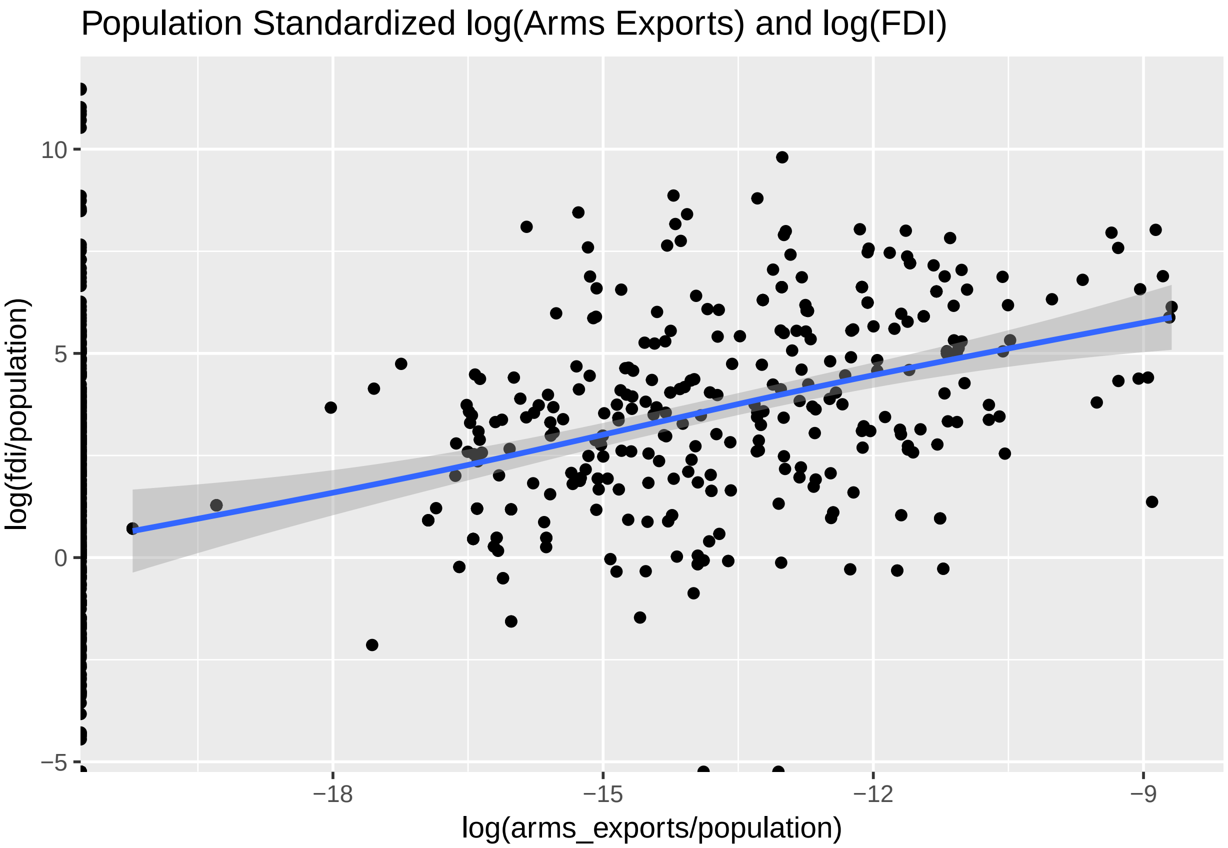
```
## Warning in log(fdi): NaNs produced

## Warning in log(fdi): NaNs produced

## Warning in log(fdi): NaNs produced

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 748 rows containing non-finite values (stat_smooth).

## Warning: Removed 244 rows containing missing values (geom_point).
```



This looks much better! Let's just try one more, dividing both values by population and then taking the log:

```
ggplot(na.omit(df), aes(x=log(arms_exports/population), y=log(fdi/population))) +
    geom_point() +
    geom_smooth() +
    ggtitle('Population Standardized log(Arms Exports) and log(FDI)')
```

```
## Warning in log(fdi/population): NaNs produced


## Warning in log(fdi/population): NaNs produced


## Warning in log(fdi/population): NaNs produced

## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'

## Warning: Removed 748 rows containing non-finite values (stat_smooth).

## Warning: Removed 244 rows containing missing values (geom_point).
```



BEAUTIFUL. This graph actually looks so good I feel like I've cheated somewhere?? Those residuals are going to be so normal.

Misc. Notes for Ben

TODO

1. Create some kind of standardized mapping of countries so that more of them get passed through all the joins.
2. Consider some strategies to fill in missing values. E.g., Polity does not assign a regime for Afghanistan for several years because of the conflict there—what makes sense as a way to handle this reasonably without just dropping the rows?
3. Fill all NAs with zeros where NA indicates absence of phenomena.
4. Attempt three models: regular `lm`, `plm` for panel data, and original two-step least square regression of the original studies, paying particular attention to showing the violation of `lm` assumptions and how the latter two correct this.

Modeling

Note on the original studies' two-stage least squares regression:

1. Authors first developed a *troop* model:

$$troops \sim conflict_{-1} + alliance_{-1} + polity_{-1} + warsaw\_pact_{-1} + cold\_war_{-1} +$$

$$log(pop_{-1}) + reagan_{-1} + south\_korea + vietnam + philippines$$

2. Then plugged the results of that model into a *trade* model with some other variables:

$$trade \sim troops + growth_{-1} + gdp_{-1} + distance + alliance_{-1}$$

where -1 subscript indicates lag of 1 year; both models including intercepts.

They do this because they suspect trade and presence of U.S. troops are endogenous to eachother; in stastical langauge there will be a correlation with model errors.