Modeling the Effect of U.S. Arms Transfers on FDI

Ben Horvath

December 12, 2018

Contents

Load libraries:

```
library(corrplot)
library(dplyr)
library(ggplot2)
library(lme4)

source('../R/multiplot.R')
```

1 Introduction

Corporations and investors can directly invest in enterprises in foreign countries, as opposed to, for instance, portfolio investments like stocks and bonds. This type of investment is called *foreign direct investment*.

Some countries will have specific advantages over others that attract FDI. Rugman (2001, 157–58) divides them into 'harder' and 'softer' advantages. The former are primarily economic—access to natural resources, a cheaper component of production, etc. Soft locational advantages refer to intangible benefits, e.g., subjective firm director preferences.

Political factors must be counted among locational advantages or disadvantages. When a left-wing government comes to power and threatens nationalization of large industries, investors are likely to become wary. International politics advantages have been well-studied. Many scholars have tested and confirmed the hypothesis that 'alliances have a direct, statistically significant, and large effect on bilateral trade' (Gowa 1994, 54; see also Gowa and Mansfield 1993 and Long 2003). Other scholars have examined FDI in this context. Biglaiser and DeRouen (2007) and Little and Leblong (2004) found that the presence of U.S. troops in a potential host country increases the level of incoming FDI, i.e., is a locational advantage to investors.

This analysis provides another test of the relationship between FDI and international security arrangements: Are U.S. arms transfers, like the presence of the U.S. military itself, a locational advantage to investors? There are a number of major reasons to believe this might be so. First, military aid suggests a friendly atmosphere between the U.S. and potential host. Second, corporate interests often correspond to the interests of the U.S. government: To quote Gilpin, "although the interests of American corporations and U.S. foreign policy objectives have collided on many occasions, a complementarity of interests has tended to exist between the corporations and the U.S. government" (1987, 241). Third, like the presence of U.S. soldiers, military aid can (but does not always) signal stability to investors and a decreased chance of political or econmomic disruption.

To test this theory, I assemble a data set of FDI flows from the U.S., military sales from the U.S., and supplementary variables likely to also affect FDI

flows. I run numerous kinds of regressions, in search of the best model of this phenomenon.

2 Data Collection

I assemble a dataset with the following variables. For full citations, see the references at the end of the paper.

- **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 is very detailed in describing exactly the kind of conflict. Instead, I will 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. Although Polity2 is a scale from -10 to 10, it is often coded to an ordinal variable with the following levels: autocracy, closed anocracy, open anocracy, democracy, full democracy, and conflict/occupied.

- **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

Assembling the complete dataset is a lengthy affair. I have contained it to the `R/clean_data.R` script.

I will simply load the finished dataset here:

```
df <- read.csv('../data/clean/master_dataset.tsv', sep='\t',
               stringsAsFactors=FALSE)
```

The primary barrier to assembling this data set was standardization of country names. For instance, in some data sets, the Vatican is called 'Holy See (Vatican City State),' and others, simply 'Holy See.' There are also occasional typos, e.g., 'NewZealand' or 'SriLanka' rather than 'New Zealand' or 'Sri Lanka.' Countries with accented characters posed another issue.

Tediously and laborously, I standardized each seperate piece of the dataset by hand.

There is one more wrinkle to deal with. Since we are testing a theory about how agents react to information, our dataset has to reflect that agent's knowledge. To do so, each of the independent variables is lagged by one year. Thus, for instance, the model will be trained on data from 2003 to predict a country's `fdi` in 2004:

```
df <- df %>%
    group_by(country) %>%
    mutate(population=lag(population, order_by=year),
           arms_exports=lag(arms_exports, order_by=year),
           conflict=lag(conflict, order_by=year),
```

```
            alliance=lag(alliance, order_by=year),
            gdp=lag(gdp, order_by=year),
            gdp_perc_growth=lag(gdp_perc_growth, order_by=year),
            regime_type=lag(regime_type, order_by=year)) %>%
    na.omit()
```

The accurately evaluate each of the models, I partition the data into separate train and test sets, where the last three years of the dataset for each country are loaded into the test set (approximately 30 percent of the observations). This will provide a good measure of how well the developed models can be expected to perform in reality.

```
train <- df %>% filter(year <= 2010)

test <- df %>% filter(year > 2010)

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

# remove from workspace for now to keep models uncontaminated
rm(test)
```

3 Exploratory Data Analysis

The training set:

```
head(train)
```

```
## # A tibble: 6 x 11
## # Groups:   country [1]
##    year country      fdi population arms_exports conflict alliance km_dist
##   <int> <chr>      <dbl>      <int>        <int>    <int>    <int>   <int>
## 1  2005 Afghani~  0.         24118979            0        1        0   11132
## 2  2006 Afghani~  0.         25070798           19        1        0   11132
## 3  2007 Afghani~  0.         25893450            0        1        0   11132
## 4  2008 Afghani~  0.         26616792           22        1        0   11132
```

```
## 5  2009 Afghani~ -1.00e6   27294031          78        1        0   11132
## 6  2010 Afghani~ -1.00e6   28004331         280        1        0   11132
## # ... with 3 more variables: gdp <dbl>, gdp_perc_growth <dbl>,
## #   regime_type <chr>
```

To get a sense of the range of our dependent variable and main independent variable:

```
mean(train$fdi)
```

```
## [1] 1488869736
```

Mean `fdi` is about $1.5 billion, while median is only $1 million. This suggests a *highly* right-skewed distribution:

```
quantile(train$fdi, c(0, .25, 0.5, 0.75, 0.9, 0.95, 0.99))
```

```
##          0%         25%         50%         75%         90%         95%
## -1.9284e+10  0.0000e+00  1.0000e+06  4.4700e+08  3.0580e+09  8.5720e+09
##         99%
##  3.0351e+10
```

The min of `fdi` is -$20 billion. Negative FDI seems puzzling. According to this World Bank explainer https://www.oecd.org/daf/inv/FDI-statistics-explanatory-notes.pdf,

> FDI financial transactions may be negative for three reasons. First, if there is disinvestment in assets— that is, the direct investor sells its interest in a direct investment enterprise to a third party or back to the direct investment enterprise. Second, if the parent borrowed money from its affiliate or if the affiliate paid off a loan from its direct investor. Third, if reinvested earnings are negative. Reinvested earnings are negative if the affiliate loses money or if the dividends paid out to the direct investor are greater than the income recorded in that period. Negative FDI positions largely result when the loans from the affiliate to its parent exceed the loans and equity capital given by the parent to the affiliate. This is most likely to occur when FDI statistics are presented by partner country.

W see that middle 50 percent of the variable is between $0 and about $150 million, and the top one percent is greater than $30 billion.

arms_exports is less extreme, but still skewed,

```
mean(train$arms_exports)
```

```
## [1] 44.59256
```

```
quantile(train$arms_exports, c(0, .25, 0.5, 0.75, 0.9, 0.95, 0.99))
```

```
##    0%   25%   50%   75%   90%   95%   99%
##   0.0   0.0   0.0   5.0  89.0 303.0 847.8
```

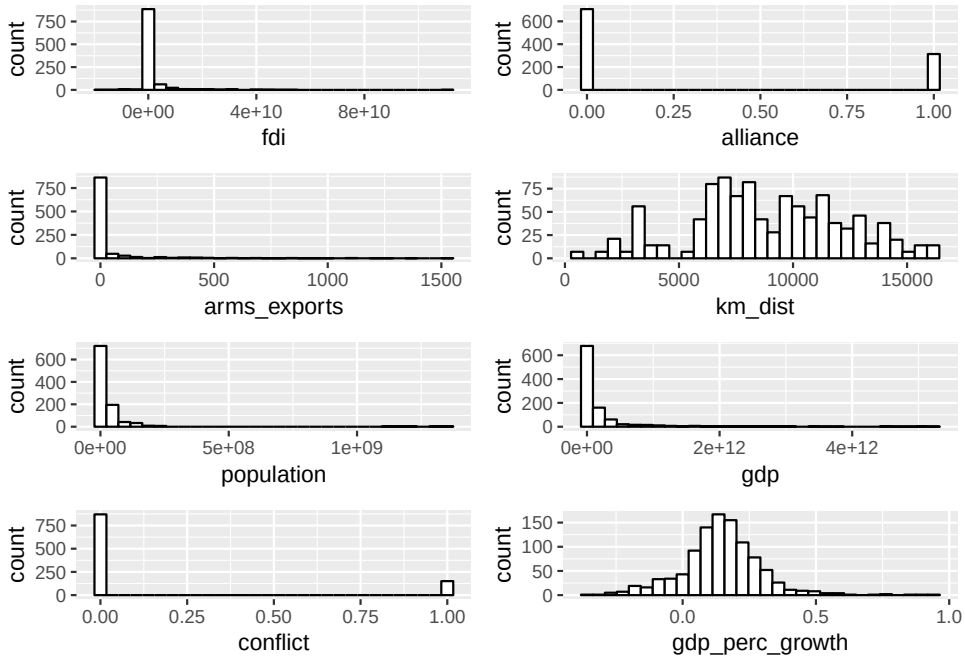
with a mean of 44.5 arms units and a median of 0. The 25th and 75th percentile ranges from 0 to 5. The largest arms transfers seem to make up about 5 percent of the total data set.

3.1 Distributions

For the first pass through, let's examine the distributions of the numeric variables:

```
hist_fdi <- ggplot(train, aes(x=fdi)) +
    geom_histogram(colour="black", fill="white")

hist_arms <- ggplot(train, aes(x=arms_exports)) +
    geom_histogram(colour="black", fill="white")

hist_pop <- ggplot(train, aes(x=population)) +
    geom_histogram(colour="black", fill="white")

hist_conflict <- ggplot(train, aes(x=conflict)) +
    geom_histogram(colour="black", fill="white")

hist_alliance <- ggplot(train, aes(x=alliance)) +
    geom_histogram(colour="black", fill="white")

hist_dist <- ggplot(train, aes(x=km_dist)) +
    geom_histogram(colour="black", fill="white")

hist_gdp <- ggplot(train, aes(x=gdp)) +
    geom_histogram(colour="black", fill="white")
```

```
hist_growth <- ggplot(train, aes(x=gdp_perc_growth)) +
    geom_histogram(colour="black", fill="white")

multiplot(hist_fdi, hist_arms, hist_pop, hist_conflict,
          hist_alliance, hist_dist, hist_gdp, hist_growth, cols=2)
```



From these graphs it's clear there are only two 'nice' variables: `km_dist` and `gdp_perc_growth`, i.e., they are approximately normally distributed.

Our two most important variables, `fdi` and `arms_exports`, are not normal. They are both *zero-inflated*, with a long right tail. This may prove challenging in attempting to model them with standard linear regression.

The variables `gdp` and `population` also have a small central tendancy, with a long right tail. This reflects the fact that most countries have a small population with a correspondingly small GDP, and that there are a few large countries with large GDPs.

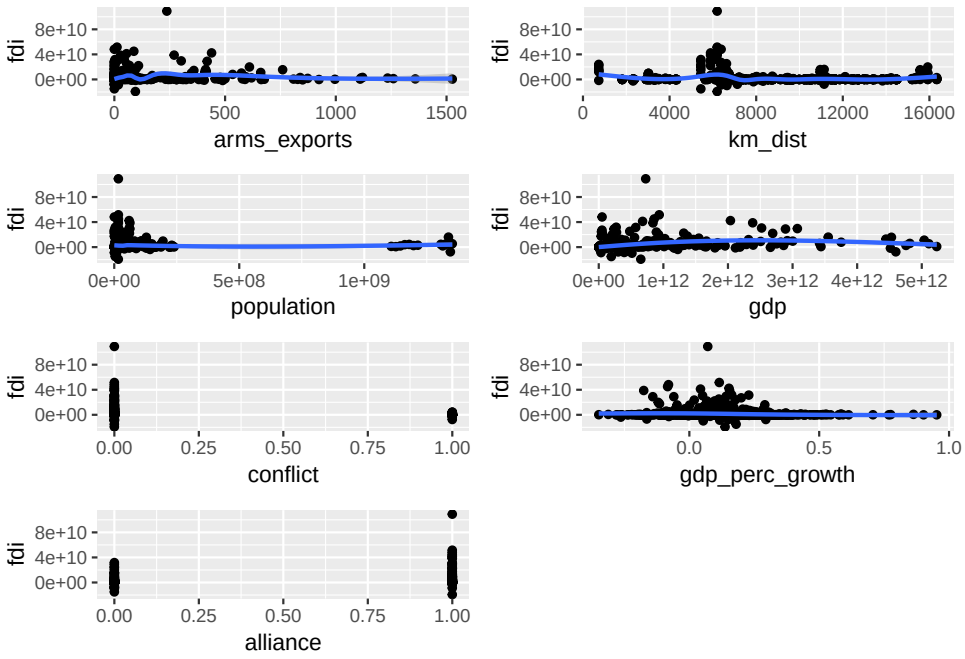
The remaining variables are dichotomous, `conflict` and `alliance`. These plots show that most of the observations in the dataset are in a time of peace, and that most of them did not occur when the country was allied

with the United States.

3.2 Associations and Correlations

The next step is to take a look for associations in our data set:

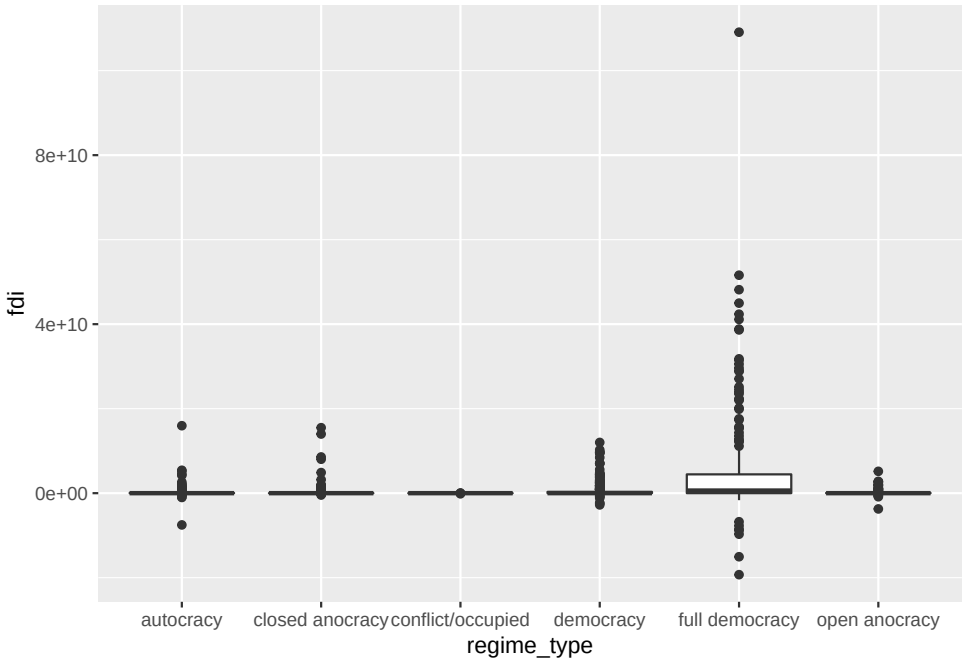
```
scat_arms <- ggplot(train, aes(x=arms_exports, y=fdi)) +
    geom_point() +
    geom_smooth()

scat_pop <- ggplot(train, aes(x=population, y=fdi)) +
    geom_point() +
    geom_smooth()

scat_conflict <- ggplot(train, aes(x=conflict, y=fdi)) +
    geom_point() +
    geom_smooth()

scat_alliance <- ggplot(train, aes(x=alliance, y=fdi)) +
    geom_point() +
    geom_smooth()

scat_dist <- ggplot(train, aes(x=km_dist, y=fdi)) +
    geom_point() +
    geom_smooth()

scat_gdp <- ggplot(train, aes(x=gdp, y=fdi)) +
    geom_point() +
    geom_smooth()

scat_growth <- ggplot(train, aes(x=gdp_perc_growth, y=fdi)) +
    geom_point() +
    geom_smooth()

multiplot(scat_arms, scat_pop, scat_conflict, scat_alliance,
          scat_dist, scat_gdp, scat_growth, cols=2)
```



Few of these variables are related to `fdi` in a straight-forward, linear way.

Examining our categorical variable, `regime_type`:

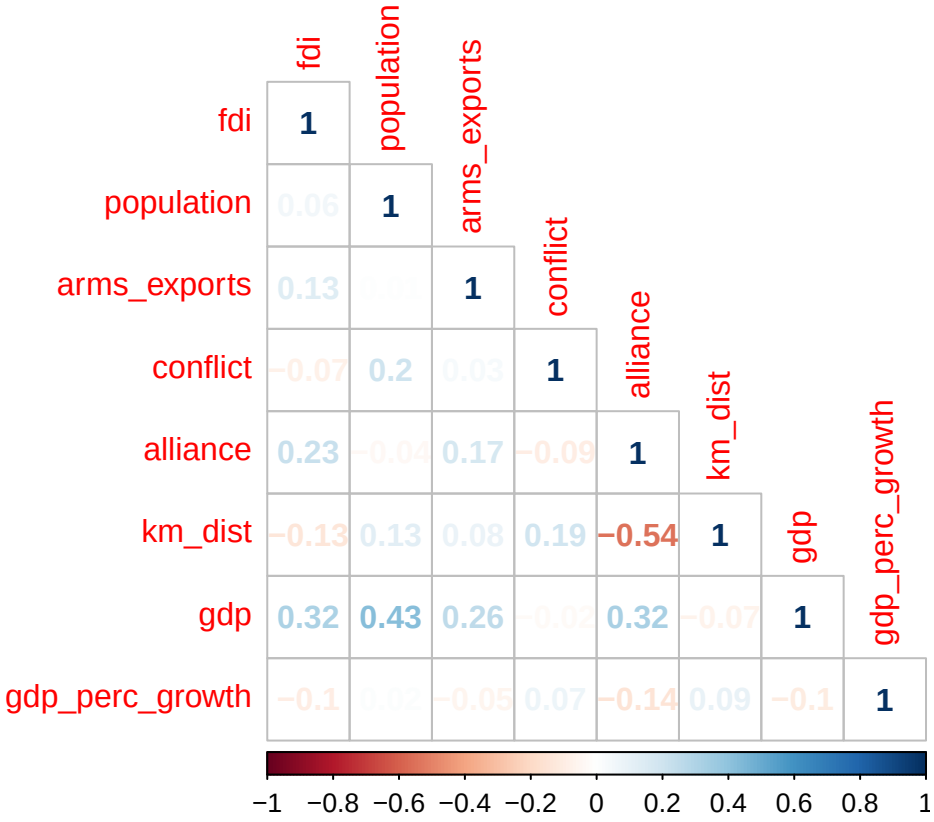
```
ggplot(train, aes(x=regime_type, y=fdi)) +
    geom_boxplot()
```



Full democracies obviously receive the most FDI from the U.S., though autocratic countries still receive some. Conflicted/occupied countries recieve very little, which makes sense.

The correlation matrix of the numerical variables:

```
train_cor <- cor(train[, c(3:10)])
corrplot(train_cor, type='lower', method='number')
```



Unfortunately for us, the correlation between the main variables of interest, `fdi` and `arms_exports` is low, only 0.13. But, more hopefully, few of our other independent variables are correlated, which protects our regression models from multicollinearity.

The two worrying relationships are `gdp` and `population`, which are naturally related, with a correlation of 0.43. More interesting is the correlation of `km_dist` and `alliance` (-0.54). This suggests that the further away a country is from the U.S., the less likely an alliance with it will be. This is a non-intuitive finding, but could possibly be explained by the difficulty of projecting force across large distances and oceans.

4 Statistical Analysis

4.1 Inferential Statistics

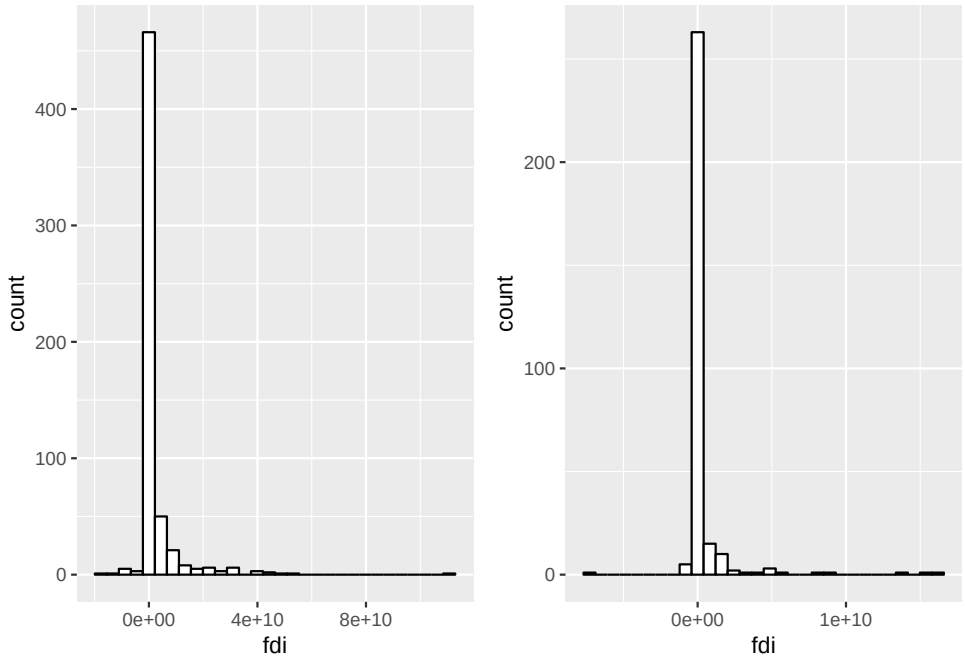
We've seen above that appears to be some kind of association between `regime_type` and `fdi`, where democracies recieve more trade that autocracies. This section will formally test this, with the hypotheses,

H_0: $\mu_{democracy} - \mu_{autocracy} = 0$

H_1: $\mu_{democracy} - \mu_{autocracy} > 0$

I create two vectors of `fdi`, one for country-years where the country was either a 'full democracy' or a 'democracy,' and the other for countries that are an 'autocracy' or 'closed anonocracy'. Their respective distributions are plotted:

```
x_demo <- train %>%
    filter(regime_type == 'full democracy' | regime_type == 'democracy') %>%
    select(fdi)

x_auto <- train %>%
    filter(regime_type == 'autocracy' | regime_type == 'closed anocracy') %>%
    select(fdi)

hist_demo <- ggplot(x_demo, aes(x=fdi)) +
    geom_histogram(colour="black", fill="white")
hist_auto <- ggplot(x_auto, aes(x=fdi)) +
    geom_histogram(colour="black", fill="white")

multiplot(hist_demo, hist_auto, cols=2)
```



The distributions suggest a couple of problems with this hypothesis test: Because of the skew and the excessive zeros, neither of these aappear close to a normal distribution. Additionally, the variance of the two samples are quite different. I will carry on the hypothesis test, with the hopes that the larger sample size and R's `var.equal=FALSE` setting will carry me through:

```
t.test(x_demo$fdi, x_auto$fdi, alternative='greater',
       var.equal=FALSE, conf.level=0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  x_demo$fdi and x_auto$fdi
## t = 5.6175, df = 686.26, p-value = 1.409e-08
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1407487063        Inf
## sample estimates:
##  mean of x  mean of y
## 2365246998  373876221
```

This t-test suggetss we should reject the null hypothesis H_0 that the mean `fdi` is the same for both democracies and autocracies. This result is very significant, with a t-value of 5.62! The larger sample size, the obvious difference between the two means, and the high significance allay most of my concerns about the violations noted above.

4.2 Models

Each model will be evaluated by its performance on the test set. The metric to optimize is means squared errors (MSE):

```
mse <- function(m) mean(resid(m)^2)

calc_r2 <- function(y, y_hat) {
    rss <- sum((y_hat - y)^2)
    tss <- sum((y - mean(y_hat))^2)
    return(1 - (rss/tss))
}
```

Attention will be paid to R^2 as well as performance on training set. However, MSE on the test set is the ultimate metric to minimize.

4.2.1 M_0: Predicting the Mean

For the purposes of establishing a baseline performance, the first model will be a dummy model, predicting only the average FDI.

```
m0 <- lm(fdi ~ 1, train)
mse(m0)
```

```
## [1] 4.024362e+19
```

With an MSE of over $4e^{19}$ (dollars), this model performs very poorly. Hopefully further iteration can improve it.

4.2.2 M_1: Linear Model, All Variables

```
m1 <- lm(fdi ~ population + arms_exports + as.factor(conflict) +
               as.factor(alliance) + km_dist + gdp + gdp_perc_growth +
               as.factor(regime_type), train)
summary(m1)
```

```
##
## Call:
## lm(formula = fdi ~ population + arms_exports + as.factor(conflict) +
##     as.factor(alliance) + km_dist + gdp + gdp_perc_growth + as.factor(regime_type),
##     data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.502e+10 -1.125e+09 -1.353e+08  3.409e+08  1.029e+11
##
## Coefficients:
##                                            Estimate Std. Error t value
## (Intercept)                                9.639e+08  8.463e+08   1.139
## population                                -4.508e-01  1.483e+00  -0.304
## arms_exports                               1.802e+06  1.260e+06   1.429
## as.factor(conflict)1                      -2.019e+08  5.622e+08  -0.359
## as.factor(alliance)1                       7.686e+08  5.530e+08   1.390
## km_dist                                   -8.739e+04  6.836e+04  -1.278
## gdp                                        2.123e-03  3.586e-04   5.922
## gdp_perc_growth                           -1.321e+09  1.266e+09  -1.044
## as.factor(regime_type)closed anocracy      4.451e+08  6.798e+08   0.655
## as.factor(regime_type)conflict/occupied    1.217e+08  1.743e+09   0.070
## as.factor(regime_type)democracy           -1.620e+08  6.054e+08  -0.268
## as.factor(regime_type)full democracy       3.182e+09  7.074e+08   4.498
## as.factor(regime_type)open anocracy        5.443e+07  7.407e+08   0.073
##                                           Pr(>|t|)
## (Intercept)                                  0.255
## population                                   0.761
## arms_exports                                 0.153
## as.factor(conflict)1                         0.720
```

```
## as.factor(alliance)1                          0.165
## km_dist                                       0.201
## gdp                                        4.36e-09 ***
## gdp_perc_growth                               0.297
## as.factor(regime_type)closed anocracy         0.513
## as.factor(regime_type)conflict/occupied       0.944
## as.factor(regime_type)democracy               0.789
## as.factor(regime_type)full democracy       7.66e-06 ***
## as.factor(regime_type)open anocracy           0.941
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.847e+09 on 1008 degrees of freedom
## Multiple R-squared:  0.1614, Adjusted R-squared:  0.1514
## F-statistic: 16.17 on 12 and 1008 DF,  p-value: < 2.2e-16
```

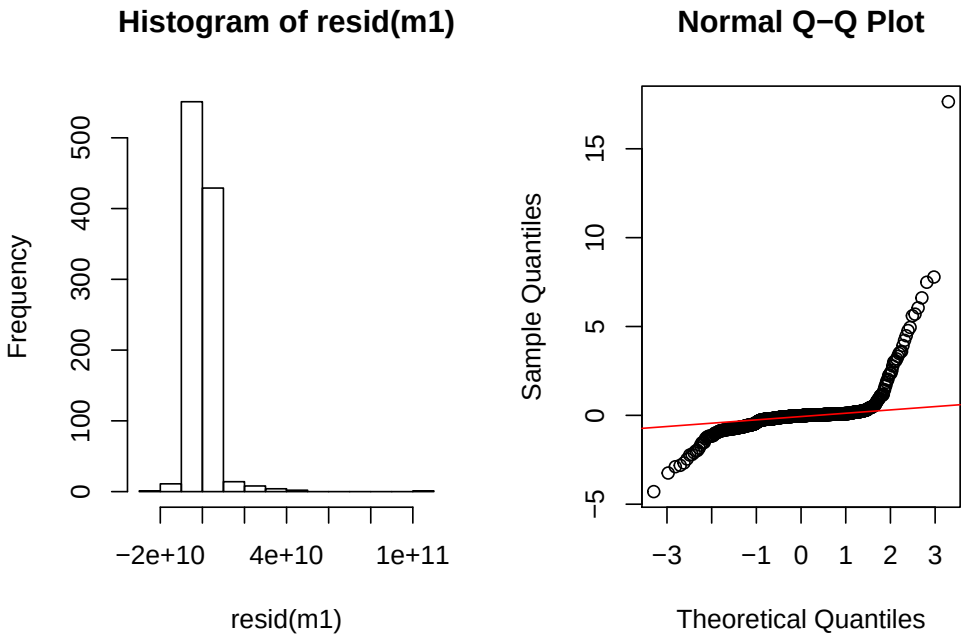
```
mse(m1)
```

```
## [1] 3.37478e+19
```

Unfortunately, this straight-forward model is not impressive. It's MSE is only about 16 percent better than predicting the average, though it does have a not-insignificant R^2 of .15, and the F-statistic says it is statistically different from the dummy model. Only GDP and regime type of full democracy are significant.

Examine the residuals:

```
par(mfrow=c(1,2))
hist(resid(m1))
qqnorm(rstandard(m1)); qqline(rstandard(m1), col = 2)
```



It is clear that the residuals are not as normal as we like. The model performs well for 'typical' observations (between the -2 and 2 quartiles), but fails for the quite a few outlying observations. Both positive and negative outliers have large residuals.

```
train_m1 <- train
train_m1$resid <- resid(m1)
train_m1 <- train_m1 %>%
    mutate(resid_abs = abs(resid)) %>%
    arrange(desc(resid))
head(train_m1[c('year', 'country', 'resid')], 10)
```

```
## # A tibble: 10 x 3
## # Groups:   country [4]
##     year country                resid
##    <int> <chr>                  <dbl>
##  1  2007 Netherlands    102855267532.
##  2  2009 Netherlands     45365348354.
##  3  2010 Luxembourg      43584466944.
##  4  2010 Netherlands     38522907898.
##  5  2006 Netherlands     35269523242.
```

```
##  6  2004 United Kingdom  33068662761.
##  7  2008 Netherlands     32645722874.
##  8  2010 United Kingdom  28689500402.
##  9  2008 Ireland         27763820382.
## 10  2004 Netherlands     26096392930.
```

Interestingly, the top cases with the most errors are all developed Western European allies of the U.S. M_1 seems to have over-estimated all of these cases, by tens of billions of dollars. Future work might try to account for this by including an variable indicating if a country is West European, or perhaps an (original) member of NATO.

Looking at the cases with negative residuals, the model seems to have especially underestimated China and Japan, especially in the period around the 2007–2008 years (during which there was an economic crisis). My intuition is that some state of affairs—an overheated world market, perhaps?—was directing excessive FDI to these countries over this time period. This suggests adding a variable to account for the state of the world market might help.

4.2.3 M_2: Linear Model, Some Logged Variables

One way to make these residuals more normal is to log some of the poorly behaved independent variables, transforming them to normality:

```
m2 <- lm(fdi ~ log(population) + arms_exports + as.factor(conflict) +
               as.factor(alliance) + km_dist + log(gdp) + gdp_perc_growth +
               as.factor(regime_type), train)
summary(m2)
```

```
##
## Call:
## lm(formula = fdi ~ log(population) + arms_exports + as.factor(conflict) +
##     as.factor(alliance) + km_dist + log(gdp) + gdp_perc_growth +
##     as.factor(regime_type), data = train)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.592e+10 -1.482e+09 -3.372e+08  7.731e+08  1.021e+11
```

```
##
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                             -1.554e+10  2.807e+09  -5.539
## log(population)                         -2.219e+08  1.993e+08  -1.113
## arms_exports                             9.975e+05  1.291e+06   0.773
## as.factor(conflict)1                    -7.189e+08  5.897e+08  -1.219
## as.factor(alliance)1                     9.931e+08  5.511e+08   1.802
## km_dist                                 -3.895e+04  6.910e+04  -0.564
## log(gdp)                                 8.270e+08  1.622e+08   5.099
## gdp_perc_growth                         -1.550e+09  1.267e+09  -1.224
## as.factor(regime_type)closed anocracy    1.115e+09  7.054e+08   1.580
## as.factor(regime_type)conflict/occupied  7.597e+08  1.750e+09   0.434
## as.factor(regime_type)democracy          6.077e+07  6.125e+08   0.099
## as.factor(regime_type)full democracy     2.740e+09  7.337e+08   3.734
## as.factor(regime_type)open anocracy      6.930e+08  7.635e+08   0.908
##                                         Pr(>|t|)
## (Intercept)                             3.89e-08 ***
## log(population)                         0.265983
## arms_exports                            0.439887
## as.factor(conflict)1                    0.223072
## as.factor(alliance)1                    0.071812 .
## km_dist                                 0.573041
## log(gdp)                                4.08e-07 ***
## gdp_perc_growth                         0.221193
## as.factor(regime_type)closed anocracy   0.114419
## as.factor(regime_type)conflict/occupied 0.664196
## as.factor(regime_type)democracy         0.920982
## as.factor(regime_type)full democracy    0.000199 ***
## as.factor(regime_type)open anocracy     0.364291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.857e+09 on 1008 degrees of freedom
## Multiple R-squared:  0.1584, Adjusted R-squared:  0.1484
## F-statistic: 15.82 on 12 and 1008 DF,  p-value: < 2.2e-16
```

```
mse(m2)
```

```
## [1] 3.386704e+19
```

Logging these variables is actually slightly worse than M_1, both in terms of MSE and R^2. In terms of variable significance, the only change is that alliance becomes significant at $p = .10$. Residuals are almost identical to previous model.

4.2.4 M_4: Mixed Effects Panel Model

This model attempts to deal with the fact that most subjects (states) are sampled from multiple times. This kind of *mixed effects* models adds a second layer of *random* effects to the usual regression model's *fixed* effects. This model will include country as a variable in an attempt to quantify specific differences due to a country that are not attributable to the independent variables.

```
m4 <- lmer(fdi ~ population + arms_exports + as.factor(conflict) +
                 as.factor(alliance) + km_dist + gdp + gdp_perc_growth +
                 as.factor(regime_type) + (1 | country), train)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(m4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## fdi ~ population + arms_exports + as.factor(conflict) + as.factor(alliance) +
##     km_dist + gdp + gdp_perc_growth + as.factor(regime_type) +
##     (1 | country)
##    Data: train
##
## REML criterion at convergence: 48034
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -13.1700  -0.0780  -0.0103   0.0418  16.6464
```

```
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  country  (Intercept) 1.595e+19 3.993e+09
##  Residual             1.837e+19 4.285e+09
## Number of obs: 1021, groups:  country, 154
##
## Fixed effects:
##                                            Estimate Std. Error t value
## (Intercept)                                8.819e+08  1.505e+09   0.586
## population                                 7.535e-01  2.735e+00   0.276
## arms_exports                               2.863e+06  1.357e+06   2.109
## as.factor(conflict)1                      -5.032e+08  6.691e+08  -0.752
## as.factor(alliance)1                       9.608e+08  1.022e+09   0.940
## km_dist                                   -7.679e+04  1.276e+05  -0.602
## gdp                                        1.515e-03  5.745e-04   2.636
## gdp_perc_growth                           -1.213e+09  9.812e+08  -1.236
## as.factor(regime_type)closed anocracy      2.189e+08  1.027e+09   0.213
## as.factor(regime_type)conflict/occupied    1.260e+08  2.480e+09   0.051
## as.factor(regime_type)democracy            2.548e+07  9.524e+08   0.027
## as.factor(regime_type)full democracy       3.106e+09  1.198e+09   2.592
## as.factor(regime_type)open anocracy        8.678e+06  1.041e+09   0.008

##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it

## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```
mse(m4)
```

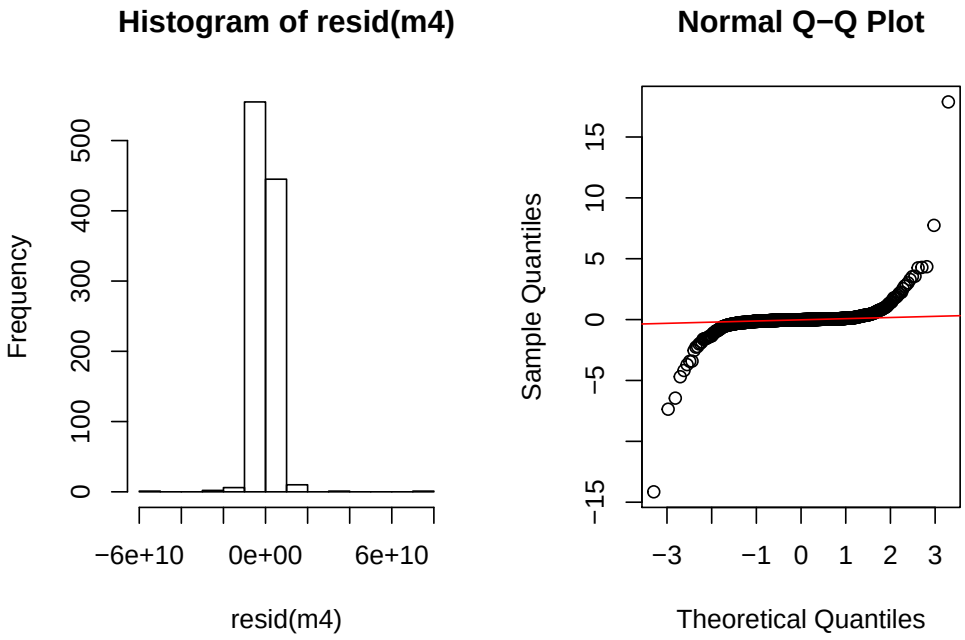
```
## [1] 1.590054e+19
```

From the MSE output, we see this model surpasses all previous models. While it has 60 percent less error than the dummy model, this is still a disappointing result.

However, with this model, `arms_exports` becomes significant at $p = 0.05$! GDP and full democracy also retain strongly significant effects.

The residual plots are also more encouraging, as many of the extreme errors we saw in M_1 disappear. The theoretical quartile plot shows a much nicer distribution, with less of a deviation from normality:

```
par(mfrow=c(1,2))
hist(resid(m4))
qqnorm(scale(resid(m4))); qqline(scale(resid(m4)), col='2')
```



Examining some of the largest residuals:

```
train_m4 <- train
train_m4$resid <- resid(m4)
train_m4 <- train_m4 %>%
    mutate(resid_abs = abs(resid)) %>%
    arrange(desc(resid))
head(train_m4[c('year', 'country')], 10)
```

```
## # A tibble: 10 x 2
## # Groups:   country [7]
##     year country
##    <int> <chr>
##  1  2007 Netherlands
```

```
##  2  2010 Luxembourg
##  3  2008 Ireland
##  4  2004 United Kingdom
##  5  2008 Switzerland
##  6  2009 Netherlands
##  7  2010 Ireland
##  8  2010 United Kingdom
##  9  2010 Australia
## 10  2004 Canada
```

Like M_1, the Netherlands and the U.K. are present, but appear less often. Other countries include Ireland, Australia, and Canada. Again, this suggests we might want to add a variable for either NATO or English-speaking countries.

4.2.5 M_5: Mixed Effects + NATO

Since it's relatively simple, let's create a dummy variable indicating whether a country was a founding member of NATO:

```
nato <- c('Belgium', 'Canada', 'Denmark', 'France', 'Iceland', 'Italy',
          'Luxembourg', 'Netherlands', 'Norway', 'Portugal',
          'United Kingdom', 'United States')

train_m5 <- train %>%
    mutate(nato=ifelse(country %in% nato, 1, 0))

m5 <- lmer(fdi ~ population + arms_exports + as.factor(conflict) +
                 as.factor(alliance) + km_dist + gdp + gdp_perc_growth +
                 as.factor(regime_type) + as.factor(nato) +  (1 | country),
           data=train_m5)
```

```
## Warning: Some predictor variables are on very different scales: consider
## rescaling
```

```
summary(m5)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
```

```
## fdi ~ population + arms_exports + as.factor(conflict) + as.factor(alliance) +
##     km_dist + gdp + gdp_perc_growth + as.factor(regime_type) +
##     as.factor(nato) + (1 | country)
##    Data: train_m5
##
## REML criterion at convergence: 47956.5
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -13.1715  -0.0696  -0.0105   0.0396  16.6870
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  country  (Intercept) 1.246e+19 3.530e+09
##  Residual             1.831e+19 4.279e+09
## Number of obs: 1021, groups:  country, 154
##
## Fixed effects:
##                                            Estimate Std. Error t value
## (Intercept)                                6.414e+08  1.370e+09   0.468
## population                                 1.521e+00  2.487e+00   0.612
## arms_exports                               3.092e+06  1.327e+06   2.329
## as.factor(conflict)1                      -4.185e+08  6.473e+08  -0.647
## as.factor(alliance)1                      -3.494e+08  9.502e+08  -0.368
## km_dist                                   -4.897e+04  1.155e+05  -0.424
## gdp                                        1.077e-03  5.429e-04   1.983
## gdp_perc_growth                           -1.233e+09  9.773e+08  -1.262
## as.factor(regime_type)closed anocracy      2.901e+08  9.638e+08   0.301
## as.factor(regime_type)conflict/occupied    8.905e+07  2.332e+09   0.038
## as.factor(regime_type)democracy            1.818e+08  8.886e+08   0.205
## as.factor(regime_type)full democracy       1.891e+09  1.125e+09   1.682
## as.factor(regime_type)open anocracy        1.828e+08  9.831e+08   0.186
## as.factor(nato)1                           9.033e+09  1.488e+09   6.071
##
## Correlation matrix not shown by default, as p = 14 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
## fit warnings:
## Some predictor variables are on very different scales: consider rescaling
```

```
mse(m5)
```

```
## [1] 1.594544e+19
```

Adding `nato` is very consequential to the model. Our main independent variable `arms_exports` becomes stronger and more significant ($p < .05$). GDP becomes less significant, though it is still significant at $p < 0.05$. Population becomes more signficiant but is 'less important.' The full democracy indicator, becomes insignicant at .05 and the magnitude of its coefficient decreases.

Interestingly, MSE is a smidge higher than M_4, by 0.2 percent. The residual graphs appear mostly the same as those of M_4, unfortunately.

5 Model Evaluations

We can now test our five models: M_0, M_1, M_2, M_4, and M_5. Reload the test data and get their predictions for the `test` set:

```
test <- read.csv('../data/clean/test.tsv', sep='\t',
                 stringsAsFactors=FALSE)

test_m0 <- test %>%
    mutate(pred = predict(m0, test),
           resid = fdi - pred)
test_m1 <- test %>%
    mutate(pred = predict(m1, test),
           resid = fdi - pred)
test_m2 <- test %>%
    mutate(pred = predict(m2, test),
           resid = fdi - pred)
test_m4 <- test %>%
    mutate(pred = predict(m4, test),
           resid = fdi - pred)

# add NATO variable in
```

```
test_m5 <- test %>%
    mutate(nato=ifelse(country %in% nato, 1, 0))
test_m5$pred <- predict(m5, test_m5)
test_m5$resid <- test_m5$fdi - test_m5$pred
```

Calculate MSE for each (divided by 10^{18} for readability), in order of best to worst:

```
paste('M_5:', mean(test_m5$resid^2) / 10^18)
```

```
## [1] "M_5: 18.3317314308561"
```

```
paste('M_4:', mean(test_m4$resid^2) / 10^18)
```

```
## [1] "M_4: 18.7364181449195"
```

```
paste('M_2:', mean(test_m2$resid^2) / 10^18)
```

```
## [1] "M_2: 51.6602441882049"
```

```
paste('M_1:', mean(test_m1$resid^2) / 10^18)
```

```
## [1] "M_1: 53.8012662506875"
```

```
paste('M_0:', mean(test_m0$resid^2) / 10^18)
```

```
## [1] "M_0: 63.9952822690553"
```

Immediately it is clear that the mixed models, M_4 and M_5, have far superior performance over the 'vanilla' M_1 and M_2 (with logged variables). Interesting, even though adding the `nato` varible slightly decreased in-sample MSE, it improved the model on the test set.

Calculate R^2, from best to worst:

```
paste('M_5:', calc_r2(test_m5$fdi, test_m5$pred))
```

```
## [1] "M_5: 0.712975771040528"
```

```
paste('M_4:', calc_r2(test_m4$fdi, test_m4$pred))
```

```
## [1] "M_4: 0.706469257997731"
```

```
paste('M_2:', calc_r2(test_m2$fdi, test_m2$pred))
```

```
## [1] "M_2: 0.188953574659956"
```

```
paste('M_1:', calc_r2(test_m1$fdi, test_m1$pred))
```

```
## [1] "M_1: 0.156506262254786"
```

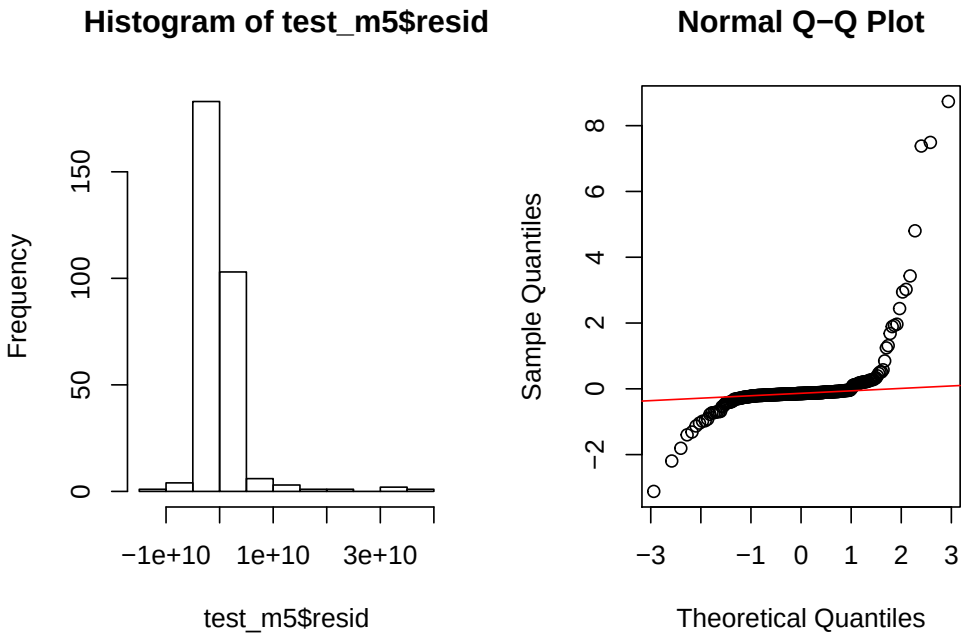
```
paste('M_0:', calc_r2(test_m0$fdi, test_m0$pred))
```

```
## [1] "M_0: 0"
```

The ordering is the same as in the case of MSE. I am pleased to see the best model explains 71 percent of the variable in FDI! (Interesting, the in-sample R^2 for M_5 is only .60.)

One final look at residuals, M_5 on the test sample:

```
par(mfrow=c(1,2))
hist(test_m5$resid)
qqnorm(scale(test_m5$resid)); qqline(scale(test_m5$resid), col='2')
```



The residuals histogram appears about the same shape as that of the training set. However, the theoretical quartile plot is not as smooth—their

are many observations where predicted value is very far from their actual value:

```
test_m5 <- arrange(test_m5, desc(resid))
head(test_m5[c('year', 'country', 'resid')], 10)
```

```
##    year        country       resid
## 1  2011    Netherlands 37699414141
## 2  2011     Luxembourg 32406834743
## 3  2011         Canada 31936016110
## 4  2012 United Kingdom 20971748405
## 5  2012     Luxembourg 15131703741
## 6  2012      Australia 13388939978
## 7  2012    Netherlands 13053412619
## 8  2012         Canada 10909575647
## 9  2012        Ireland  8886915986
## 10 2012    Switzerland  8734211652
```

Among the largest residuals are the same old culprites: Netherlands, the U.K., etc.

6 Conclusion

This paper confirmed the relationship between U.S. arms sales and U.S. FDI. The more arms a country recieves from the U.S. in year t, the more direct foreign investment the country will recieve from the U.S. the next year $t + 1$. This relationship is statistically significant, even when controlling for other factors well-known to influence FDI.

I tested a number of models and found that mixed effect models best capture the data set. The final model M_5 performed the best, explaining 71 percent of variation in `fdi` on the test dataset.

Other significant predictors of FDI include GDP and NATO membership, which both have a positive effect ($p < .05$). Regime type of full democracy also has a positive relationship with FDI at a lower significance ($p < .10$).

Future work should focus on finding an explanation for why every substantial model tended to overestimated FDI in a handful of highly developed

NATO allies, especially the Netherlands and the U.K. Introducing the NATO membership as a variable helped, but was insufficient to fully account for it.

7 References

Biglaiser, Glen, and Karl DeRouen, Jr. "Following the Flag: Troop Deployment and U.S. Foreign Direct Investment." *International Studies Quarterly* 51 (4): 835-854.

Center for Systemic Peace. 2017. *Polity IV Annual Time-Series, 1800-2017* (Excel file). <http://www.systemicpeace.org/inscrdata.html>.

Gibler, Douglas M. 2013. "Formal Alliances (v4.1)." *International Military Alliances, 1648-2008*. CQ Press. <http://www.correlatesofwar.org/data-sets/formal-alliances>.

Gilpin, Robert. 1987. *The Political Economy of International Relations*. Princeton: Princeton University Press.

Gleditsch, Kristian Skrede. _Distance Between Capital Cities." <http://ksgleditsch.com/data-5.html>.

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg, and Håvard Strand. 2002. "Armed Conflict 1946-2001: A New Dataset." *Journal of Peace Research* 39 (5).

Gowa, Joanne. 1994. *Allies, Adversaries, and International Trade*. Princeton: Princeton University Press.

Gowa, Joanne, and Edward D. Mansfield. 1993. "Power Politics and International Trade." *American Political Science Review* 87 (2): 408-20.

Little, Andrea, and David Leblang. 2004. "Military Securities: Financial Flows and the Deployment of U.S. Troops." In *Annual Meeting of the American Political Science Association*. Chicago, IL.

Long, Andrew G. 2003. "Defense Pacts and International Trade." *Journal of Peace Research* 40 (5): 537–52.

OECD. 2018. "Benchmark definition, 3rd edition (BMD3): Foreign direct investment: flows by partner country." *OECD International Direct Investment Statistics* (database).

Rugman, Alan M., and Alain Verbeke. 2001. "Location, Competitiveness, and the Multinational Enterprise." In *Oxford Handbook of International Business*, ed. A. M. Rugman and T. L. Brewer. Oxford: Oxford University Press.

Stockholm International Peace Research Institute. *Arms Transfers Database*. <https://www.sipri.org/databases/armstransfers>.

World Bank. 2018. *National Accounts Data*. <https://data.worldbank.org/indicator/NY.GDP.MKTP.CD>.

United Nations, Population Division. "Total Population - Both Sexes" (Excel file). <https://population.un.org/wpp/Download/Standard/Population/>.