The Geography of China's Foreign Investment: A Data-driven Analysis of China's New Spheres of Influence

John Ferguson, Alexander Arber, Matt Kind

May 16, 2020

Introduction

Intro paragraph(s) here:

- What is your research question?

Our research question is: Is Chinese foreign investment disproportionately allocated to the birth regions of Asian foreign leaders, and if so how do the effects vary based on regime type?

A recent 2019 paper co-authored by a group of German economists partnered with the AidData research lab at the College of William & Mary found that political leaders' birth regions do, in fact receive more Chinese aid (which they term as birth-region bias) when studying the subnational allocation of Chinese aid in Africa. The paper can be found here: https://www.sciencedirect.com/science/article/abs/pii/S030438781831099X?via%3Dihub. We used a similar research method in addition to the same geocoded Chinese aid data to see if this finding was applicable to other world regions, selecting Asia as our geographic area of focus. We then used the Democracy Index from the Economist Intelligence Unit to see how disproportionate suballocation might vary based on regime type.

- Why do we care/why is it interesting?

We believe this question is interesting because findings would help to provide clarity and offer context for how to evaluate debates surrounding current lending patterns of the Belt and Road Initiative (BRI), China's enormous geostrategic infrastructure plan allocating estimates between $1 and $8 trillion US dollars across Eurasia. The geocoded Chinese aid data we study in this paper is for investments made between 2000-2014, prior to the announcement of the BRI in 2013. Thus, our findings could explain whether debate over the cultivation of elite patronage networks in an effort to build political legitimacy for foreign leaders inside BRI countries is empirically justified or even precedes the commencement of BRI.

- What are your hypotheses?

Our initial hypothesis is that there will be far less correlation in Asia than in Africa for several reasons, but for the countries where there is disproportionate allocation, we expect to see strong correlation with authoritarian countries receiving more disproportionate investment while democratic countries should theoretically receive less. Asia as a region is far more diverse in nearly every category from regime type to GDP per capita than Africa and thus Chinese companies would find it difficult to in a sense, "collude" with local authorities on projects that support incumbent leaders. Africa traditionally has had far more strongmen rulers through the modern era than in Asia in addition to a greater share of civil war, coups, military dictatorships, lower standards of living, more natural resources like oil open to exploitation (the resource curse), and higher levels of corruption which could all help explain why the German research team found there to be strong trends of disproportionate allocation.

- What method will you use to test these hypotheses?

In order to test whether leaders' birthplaces receive a larger allocation of Chinese aid, we estimate by employing a simple linear regression (bivariate regression) and multivariate regression.

bivariate regression equation: yi = α + βxi + ϵi

Bivariate Regression Equation: yi = The outcome variable α = Intercept β = Slope coefficient xi = Independent/explanatory variable ϵi = Error term

We can interpret these variables to make more sense of them. The intercept (α) is the average value of Y when X is 0. The slope (β) is the average change in Y when X increases by unit. However, we don't know the actual values of α and β because they are unknown features of the data-generating process. The error term (ϵ) makes it impossible to observe them directly. Therefore, we relied on the α and β estimates, which is a function of the data that is our best guess about the parameters.

Our multivariate regression follows a similar methodology but accounts for 6 covariates.

multivariate regression equation: Y = α + β 1x1 + β 2x2 +. . . + β pxp + ϵ

Multivariate Regression Equation: Y = The dependent variable of the regression α = Intercept β = Slope of the regression X1 = First independent variable of the regression X2 = Second independent variable of the regression X3 = Third independent variable of the regression ϵ = Error term

In this model,α is the intercept, βj represents the increase in the average outcome associated with a one-unit increase in Xj while holding the other variables constant ϵ is an error term, and p is the number of predictors and can be greater than 1. Therefore, the multivariate regression predictors enable us to assess the impact of each predictor. Similar to the linear regression model, we rely on estimates of the data to be out best guess about the parameters.

Data Section

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   row = col_double(),
##   year = col_double(),
##   Merged_row = col_double(),
##   Merged_years = col_double(),
##   leader_start_year = col_double(),
##   leader_end_year = col_double(),
##   match = col_double(),
##   project_id = col_double(),
##   latitude = col_double(),
##   longitude = col_double(),
##   location_class = col_double(),
##   geographic_exactness = col_double(),
##   transactions_start_year = col_double(),
##   transactions_end_year = col_double(),
##   project_total_commitments = col_double(),
##   interest_rate = col_double(),
##   maturity = col_double(),
##   grace_period = col_double(),
##   grant_element = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
## [1] 19
```

Data explanation here:

- What dataset(s) are you using?

We aggregated a number of different sources to compile our dataset(s).

The first is a dataset of geocoded Chinese aid projects from 2000-2014 found here: https://www.aiddata.org/data/geocoded-chinese-global-official-finance-dataset. We extracted the most useful columns: subregion (dispersed across three different geographic variables), the year the project was agreed upon, and the investment amount in USD (millions) indexed to 2014 rates to account for inflation. Asia as a region includes 38 countries in Southeast Asia, East Asia, Central Asia, and the Middle East (West Asia). Small island nations or city-states like the Maldives and Singapore were excluded. Russia and Cyprus were considered European but included for geographic purposes. Countries with disputed recognition like the Palestinian Territories and Israel were both included. Countries with no investments from China included Timor-Leste, Azerbaijan, and Armenia. Because we are only examining subnational geographic-specific directed investment, this excludes national-level discretionary blanket investment that is then allocated across the country without specific suballocation. There are some financing projects which are committed to a group of countries; these were excluded. The total number ended up being 1,776 projects dispersed across 38 countries.

The second is a dataset that our team personally compiled: the birthplaces and years of rule for Asian foreign leaders in the time range of 2000-2014. Because in some cases, many countries had transitionary or interim governments, short-lived or alternating military dictatorships (Pakistan), or leadership disputes (Myanmar), we made a decision to simplify the scope of our project and just select the top 3 longest serving leaders in our time range. There may not necessarily be 3 leaders per country; that was just our limit: many countries just had one or two leaders that covered the range of 2000-2014. In defining tenure, we took the longest in terms of raw years (ignoring months or days), so it may be the case that there would be a head of state who actually served longer, but only marginally. We also considered that across the number of diverse regime types in Asia, in some cases we had a number of government structures with different leadership positions with powers that varied widely i.e. President, Prime Minister, King, Sultan, Supreme Leader, General Secretary, Ayatollah, etc. For consistency, we chose to go with "leaders whose offices constitutionally administer the executive or legislature of their respective state/government" according to these statistics: https://en.wikipedia.org/wiki/List_of_current_heads_of_state_and_government. A number of leaders were born on foreign soil, typically in neighboring countries or in Western nations, which were thus excluded. When it came to countries with multiple recognized governments like Yemen, we went with the official UN-recognized government. For leaders whose birth region is disputed such as Kim Jong-Il, we went with international accounts of speculation. Some leaders were born in areas that are not current countries, but former colonial imperial possessions (the former USSR the most salient example); these are coded as their modern-day equivalent nation-state. Some leaders have served multiple, nonconsecutive terms (Vladimir Putin) but they are counted as separate leaders in our account provided there is different tenure.

Merging these two datasets: we appended two variables of the second dataset to the first by matching every investment project via year and country to a) who the incumbent foreign leader was at the time of the project being signed and b) where they were born. This was done using Microsoft Excel, a Microsoft Access server database, and an ASP script that looped through to create these matches. We then rexported to Excel, then converted to CSV and uploaded to R as the "Merged" dataframe. Once every project had a leader and an associated birth region, we then went through each country and paired geographic matches between the project subregion and the birth region. While we could have used text match algorithms, this was technically unfeasible given the high prevelance of special characters from foreign languages which would have made the task impossible.

The third is another dataset that our team personally compiled: a number of aggregate country statistics, including our independent variable along with a number of possible confounders. These are found in the "countryIndicatorsDF" dataframe.

```
+ What is the size of your sample?
+ What is the unit of analysis?
+ What is your research design (cross sectional, randomized experiment, etc)
```

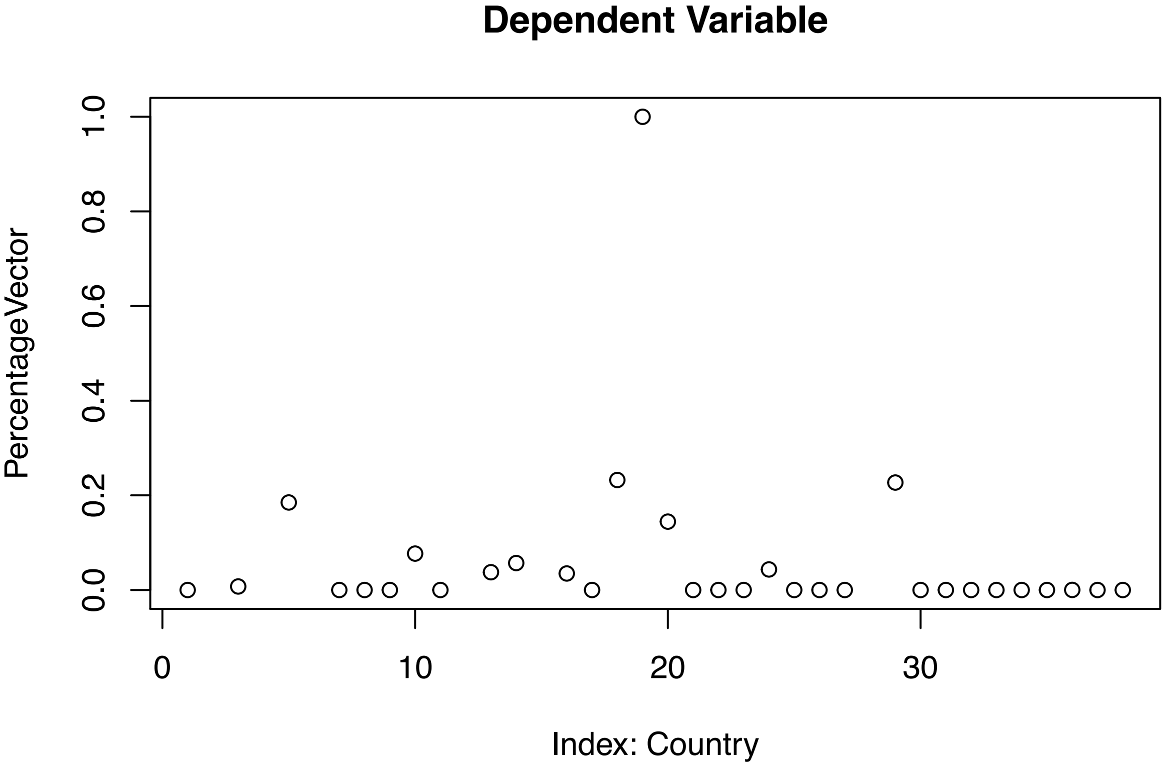
Thus our final size of the sample was 38 countries. Our unit of analysis was the country. Our research design was cross sectional.

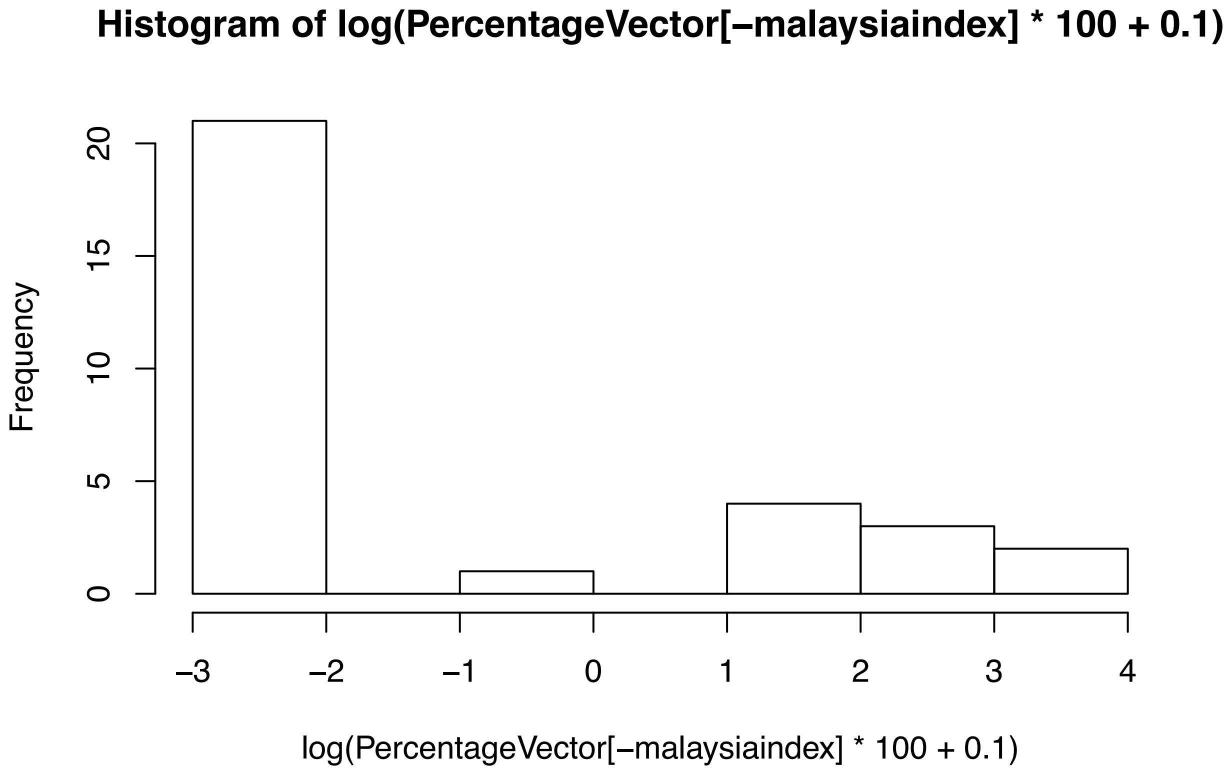
- What is your dependent variable and how is it coded? If you cleaned your data in some way, how did you do it? (i.e. recoded data from percent to decimal, dropped some irrelevant responses, etc.)

We constructed our dependent variable into a single number per country: a ratio of investment toward birth regions compared to investment in the whole country. We did this by first creating a binary variable that represents if there is a match with the location of the project and the birth area of the foreign leader. This was created into its own subset. A loop went through each country, creating temporary datasets for total projects and matched projects, multiplying them by their investment dollar amounts, and outputting the final ratio.

We cleaned our dependent variable by removing Malaysia, which was an extreme outlier as well as log transforming our dependent variable.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
## 0.00000 0.00000 0.00000 0.06393 0.03900 0.99987       6 
```

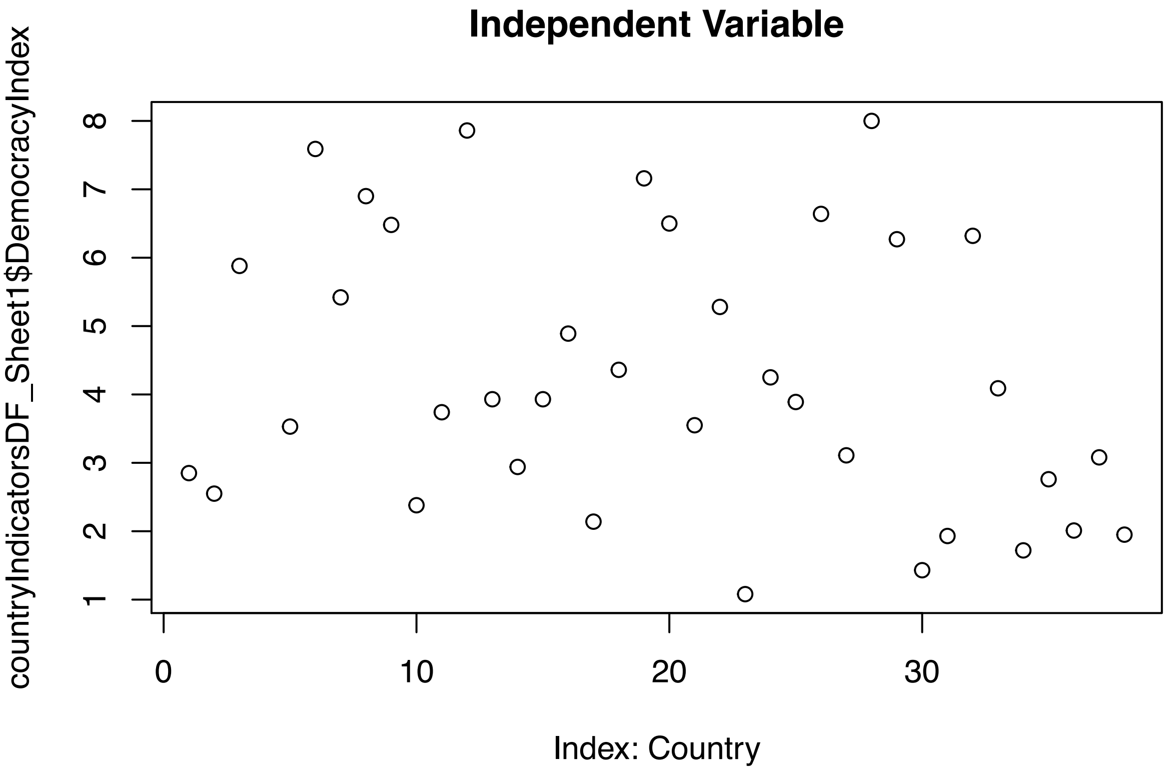




- What is your independent variable and how is it coded? If you cleaned your data in some way, how did you do it? (i.e.. recoded data from percent to decimal, dropped some irrelevant responses, etc.)

Our indpendent variable is the democracy index, a weighted average index based on 60 indicators with expert assessment that describes how democratic a country is on a scale of 1-10, compiledby the Economist Intelligence Unit. We found our data here: https://en.wikipedia.org/wiki/Democracy_Index.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##   1.080   2.760   3.930   4.281   6.270   8.000       1
```



- Are you considering any confounders? How are they coded?

We considered several confounders that could impact our results and thus collected a number of aggregate statistics for each country. We considered these statistics to be confounders and are coded in the following way:

PPPGDPPerCapita - GDP Per Capita Measured in PPP (Purchasing Price Parity): This statistic is a measure of per capita wealth adjusted for the relative cost of living and inflation rate per country, thus more useful than GDP. It could be the case that more democratic countries are naturally just wealthier countries. We used statistics from: https://en.wikipedia.org/wiki/List_of_countries_by_GDP_(PPP)_per_capita.

RegionNumber - Number of first-level administrative divisions: These first-level administrative regions are how the AidData is categorized as to the suballocation of the investment. Naturally, countries with less administrative regions will likely receive more matches just due to a smaller number of possible areas where investment could go. We used statistics from: https://en.wikipedia.org/wiki/List_of_administrative_divisions_by_country#Administrative_divisions_with_ISO_3166-1.

UrbanAreasOver1Mil - Urban Areas Over 1 Million People: In many countries, especially poorer or smaller ones, economic activity and thus foreign investment tends to cluster in one or two major economic hubs with larger populations. This could confound the result as a foreign leader could be born in these economic hubs (most likely the capital), but it would be misleading to read disproportionate investment as political bias given that there are just simply no other alternatives. We used statistics from: https://en.wikipedia.org/wiki/Number_of_urban_areas_by_country.

TotalLandArea - Total Land Area: Many countries are just simply too small too support more investment projects to build infrastructure or hold urban centers to support populations. We used statistics from: https://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_area.

DebtasPercentofGDP - Total Debt as a Percentage of Total GDP: This statistic helps to account for relative amounts of debt. This is also referred to as national debt, and could influence the likelihood that a country receives a loan. We thought that this metric could be a confounder as China could choose to increase its investment in a certain country if they have a higher debt ratio (in other words, they are more willing / likely to take on more debt without extensive consideration). We used statistics from: https://worldpopulationreview.com/countries/debt-to-gdp-ratio-by-country/.

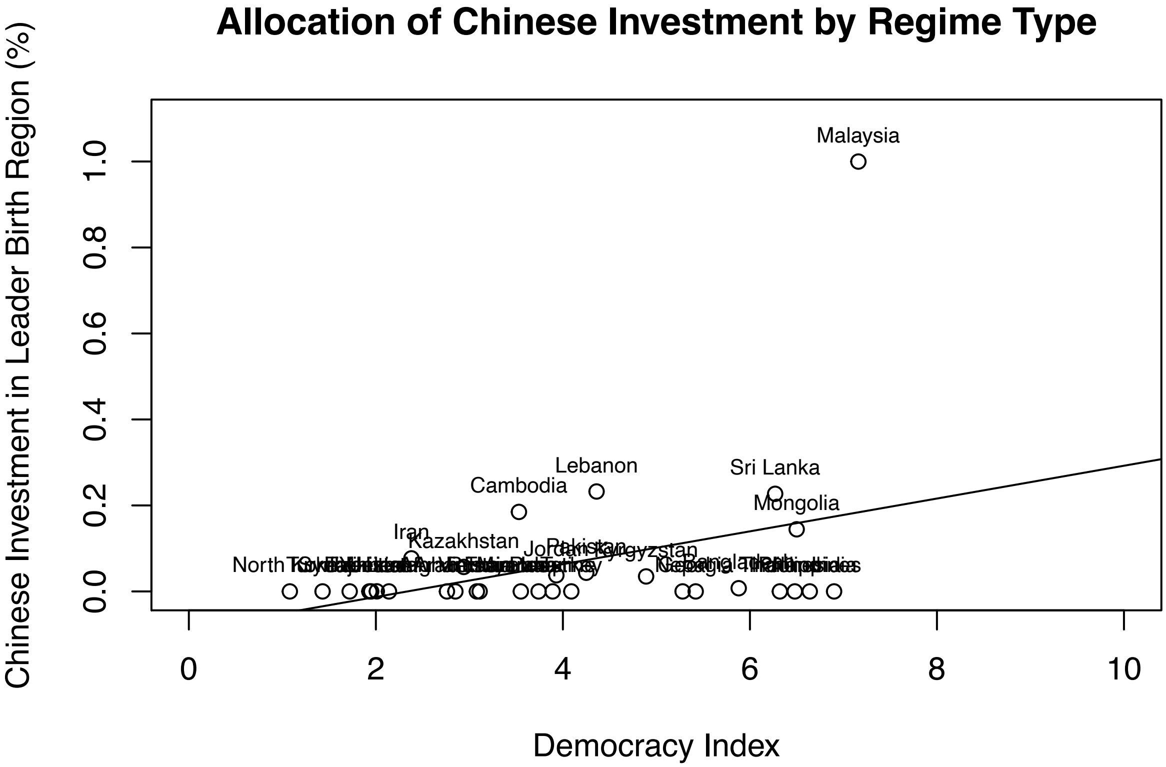
Gini - Gini Coefficient: While GDP per capita is the measure of wealth which we included, including the gini coefficient, a measure of income inequality helps to offset the confounding of extreme wealth concentrated in political elites. We used statistics from: https://worldpopulationreview.com/countries/gini-coefficient-by-country/.

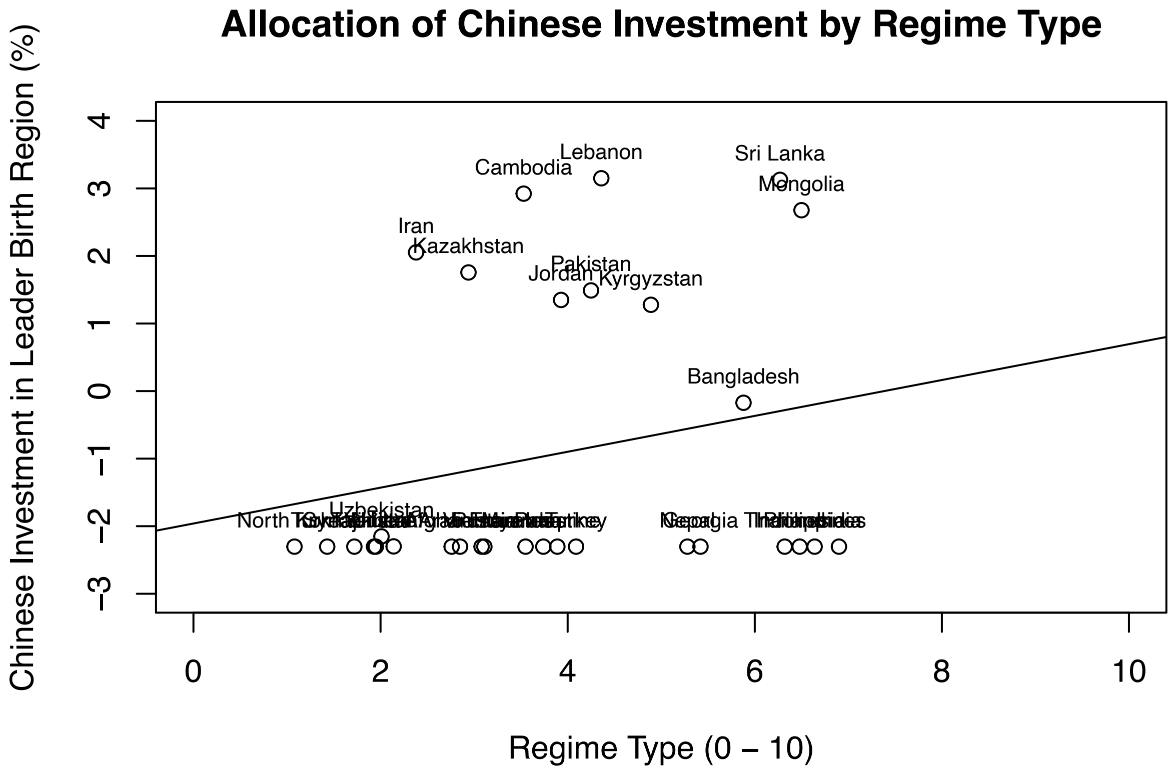
Results

Your results paragraphs should answer the following questions:

- What is the observed relationship when you plot it?

We have made two graphs here: one with Malaysia and not log-scaled and one where we took out Malaysia and log-scaled the graph. Malaysia was a clear outlier that we didn't want to significantly affect the results. Viewing the graph(s), it appears that there is a positive correlation between our dependent and independent variable. With Malaysia—it looks like we get a counterintuitive result—that more democratic countries (Sri Lanka, Mongolia, Malaysia) receive more aid to the birth regions while more autocratic countries actually receive less (Iran, Cambodia, Lebanon). But after taking Malaysia out, we find that really it was just Malaysia driving that result. For example, both Iran and Mongolia, which happen to be far away in terms of polity score, are being given roughly the same amount of money to the birthplaces of their leaders. But, this is descriptive because we aren't controlling for anything. If we control for confounders and more beyond just a visual analysis, the relationship may become more clear.





- If using linear model: what are your bivariate regression results? Interpret your coefficient of interest, and comment on the statistical significance. Do you think your result represents a causal effect?

With a coefficient of interest value of 0.2655, this means that on average across all countries, a one-unit increase on the democratic index corresponds with a 0.2655% increase in log-transformed investment toward birth regions.

On the statistical significance, approx. 4.8% of the change in percentagevector can be explained by variation in the democracyindex. This is extremely insignificant.

No, this result is clearly not a causal effect.

```
##
## Call:
## lm(formula = log(PercentageVector[-malaysiaindex] * 100 + 0.1) ~
##     countryIndicatorsDF_Sheet1$DemocracyIndex[-malaysiaindex])
##
## Coefficients:
##                                                (Intercept)
##                                                    -1.9604
## countryIndicatorsDF_Sheet1$DemocracyIndex[-malaysiaindex]
##                                                     0.2655
```

```
## [1] 0.04845881
```

```
##
## Call:
## lm(formula = log(PercentageVector[-malaysiaindex] * 100 + 0.1) ~
##     countryIndicatorsDF_Sheet1$DemocracyIndex[-malaysiaindex])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1740 -1.3550 -0.8599  2.1029  3.9531
##
## Coefficients:
##                                                          Estimate Std. Error
## (Intercept)                                               -1.9604     0.9334
## countryIndicatorsDF_Sheet1$DemocracyIndex[-malaysiaindex]   0.2655     0.2185
##                                                          t value Pr(>|t|)
## (Intercept)                                               -2.100   0.0445 *
## countryIndicatorsDF_Sheet1$DemocracyIndex[-malaysiaindex]   1.215   0.2341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.087 on 29 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.04846,    Adjusted R-squared:  0.01565
## F-statistic: 1.477 on 1 and 29 DF,  p-value: 0.2341
```

- Do your results change when you add confounders? Interpret your coefficient of interest, and comment on the statistical significance. Do you think your result represents a causal effect?

Our results do not change when we add confounders as the coefficients on the confounding variable are extremely small. While some are slightly positive and some are slightly negative, these numbers are so infinitesimally small that it doesn't change our results. On the statistical significance, when we use a multivariate regression with an adjusted R^2, the value decreases even further to 1.26% of the variation in the model that can be explained by the multivariate model. The reason we used adjusted R^2 is because it accounts for the effect of overfitting when you have more/many additional covariates. We still do not think that our results represent any causal effect. This multivariate regression further confirms this.

```
##
## Call:
## lm(formula = PercentageVectorField ~ DemocracyIndex + PPPGDPPerCapita +
##     RegionNumber + UrbanAreasOver1Mil + DebtasPercentofGDP +
##     Gini + TotalLandArea, data = withoutMalaysiaIndex, na.action = na.exclude)
##
## Coefficients:
##        (Intercept)      DemocracyIndex     PPPGDPPerCapita        RegionNumber
##         -2.877e-02           1.010e-02           2.818e-06          -1.146e-03
## UrbanAreasOver1Mil  DebtasPercentofGDP                Gini       TotalLandArea
##         -6.610e-04           2.766e-04           8.945e-05          -2.797e-09
```

```
## [1] 0.01265076
```

Conclusion

- Summarize your findings. Is your hypothesis supported by your results?

Our findings have revealed that a relationship likely does not exist between a country's regime type (democracy index) and the amount of investment allocated by China to the birth regions of its leaders. In other words, our hypothesis is wrong as we did not find any correlation. This means that the correlation found by the German researchers on Chinese investment in Africa is not generalizable to Asia. We expect this is due to the fact that compared to Africa, regime types and governance in Asia varies dramatically more than in Africa. Additionally, Chinese investment would theoretically seek to achieve different things: it's very possible that China is seeking to cultivate relationships with African leaders to build long-term energy / natural resource relationships while that dynamic doesn't exist with Asian countries. Thus, we hypothesize that China tailors or adjusts its investment strategy according to geographic region. Because we limited our analysis to Asia, our findings can't be generalized to other geographic regions (Europe, Latin America, Oceania, etc.) and thus research into other areas of the world would be a step forward in confirming this theory. So although our hypothesis was ultimately incorrect, it still provided valuable insight into Chinese investment patterns that suggest further areas of research.

- What are some limitations to your analysis? Do you think you have identified a causal effect? Why or why not?

No, we have not identified a causal effect. Part of this may be attributed to some fundamental limitations in our analysis. First, the confounders we account for are primarily economic confounders and don't account for the diverse cultural differences between different countries in Asia (i.e. Southeast Asia is very different from the Middle East). These cultural differences could significantly affect where political elites are born in their country for example. Second, a lot of our analysis is heavily dependent on how countries define their political administrative unit (which we use to match regions). It could very much be the case that in some countries, these subregions are vast and cover disproportionate areas of the countries. In other cases, the urban-rural divide makes it difficult for birth-regions to be clear in whether they are in villages or in urban areas. Third, the usefulness of the democracy index could be debated. There is an existing literature of criticism of such rankings as this which primarily rely on expert opinion. As the Economist is also a Western publication, many of the experts could be assumed to be Western and thus be biased in their political-ideological views on such subjective matters which are open to interpretation about freedom, corruption, etc.

- How can your analysis be improved if you had access to unlimited data?

If we had access to unlimited data, our group would have liked to study investments post-2013 after the Belt and Road Initiative significantly increased the amount of outbound Chinese investment. Unfortunately after inquiring with analysts at both the Mercator Institute and CSIS (Center for Strategic and International Studies), that data was not available for the purposes of our research. Unlimited data would have plugged certain holes in our data (transactions missing, incomplete investment amounts, resolved ambiguity about administrative subregions within countries, etc.). Unlimited data would allow us to also capture what investments are made in the informal economy and not reported explicitly. Obviously if China's intent was to cultivate insider patronage elite political networks through investment as it has done in Africa, it most likely would be more secretive about where its money goes.

Individual contributions

- Please note each team member's contribution to the final product here.

All three of us contributed equally in all stages of the project. This includes collecting and refining our data, working through R, and brainstorming about design. All of three of us have been meeting weekly to step through R together and we all equally consulted with Professor Olivella and Jacob (who were both very helpful) on several occasions to get their expertise in solving certain problems.