A Look at Portfolio

Youngsoo Baek

11/14/2017

Since I came to know more about CapitalG and what it does as a company, I thought it will be a good idea to have a look at some of the statistics of the companies in CapitalG portfolio. This is a summary report of the most basic exploratory data analysis I did over two days. The ultimate goal is to address possible patterns in CapitalG's past investments and thereby construct a model that predicts future investment behaviors. Please be aware that the data I am using is manually collected from available open web source. The exact valuations of the companies are therefore nowhere near precise.

I have manually collected the data set from Crunchbase and each company's website. In case of public companies, their approximate values are computed from the balance sheets on NASDAQ website. It is saved in the following `csv` file that I have imported from MS Excel.

```
## portfolio2.csv
```

This data set includes 29 companies on the CapitalG portfolio with the following variables:

- **Company**
- **Industry**
- **Location**
- **Value**: Approximate company valuation figure at the time of initial investment
- **FoundedYear**
- **InvestYear**
- **InvestMonth**
- **InvestDay**
- **Type**: Series/type of funding round
- **Stage**: Early (before Series C, or other forms of early funding roudns); Late
- **Lead**: Whether CapitalG was a lead investor
- **Fund**: Amount of funding raised in the round of CapitalG initial investment
- **Acquisition**: The number of acquisitions made by the company up before initial investment
- **Current**: Approximate current valuation figure
- **After**: Number of funding rounds CapitalG invested in after initial investment
- **AcquiredYear**: The year the company was acquired (if acquired)
- **AcquiredMonth**
- **IPOYear**: The year the company filed for IPO (if it went public)
- **IPOMonth**

This analysis was done using `R`. The following packages are needed for reproducing the diagrams using the code:

```
library(dplyr)
library(ggplot2)
library(tibble)
library(magrittr)
```

Summary statistics

First, a cursory look at the distribution across industries:

```
## # A tibble: 12 x 2
##    Industry      count
```

```
##    <fct>         <int>
##  1 Commerce          1
##  2 Education         3
##  3 Finance           3
##  4 Gaming            1
##  5 Health            3
##  6 Manufacturing     1
##  7 Mobile app        2
##  8 Real estate       2
##  9 Security          4
## 10 Service           3
## 11 Social media      1
## 12 Software          5
```

Another summary based on the companies' locations:

```
## # A tibble: 8 x 2
##   Location count
##   <fct>    <int>
## 1 CA          17
## 2 China        1
## 3 GA           1
## 4 India        4
## 5 MA           1
## 6 NY           3
## 7 PA           1
## 8 WI           1
```

The proportions of early and late stage investments:

```
## # A tibble: 3 x 2
##   Stage percentage
##   <fct>      <dbl>
## 1 Early      20.7
## 2 Late       72.4
## 3 <NA>        6.90
```

The proportions of lead and non-lead investments:

```
## # A tibble: 3 x 2
##   Lead  percentage
##   <fct>      <dbl>
## 1 No          13.8
## 2 Yes         69.0
## 3 <NA>        17.2
```

The two figures reproduced below show the distribution of CapitalG investments over years, and the distribution of how old each company is. Since over the majority of companies are located in CA, the regions of each company are grouped into three categories: CA, India, or others.

The two companies that were founded before 1990s are Renaissance Learning and Multiplan. More figures illustrating the distribution of companies across different regions and industries are attached at the end of this document.

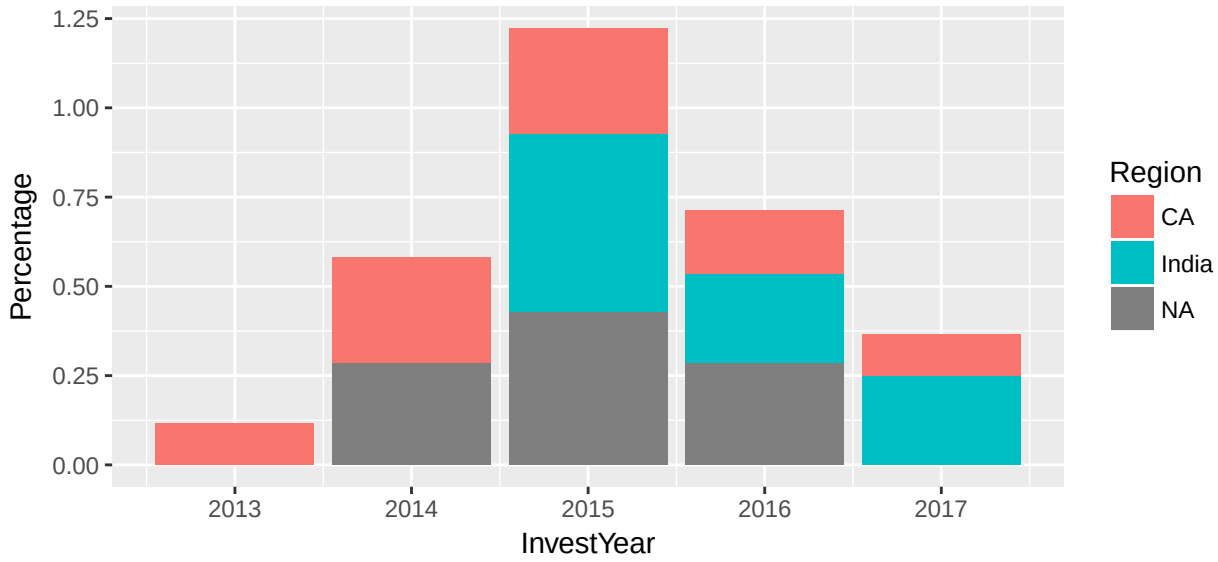

Figure 1: Investments over Years

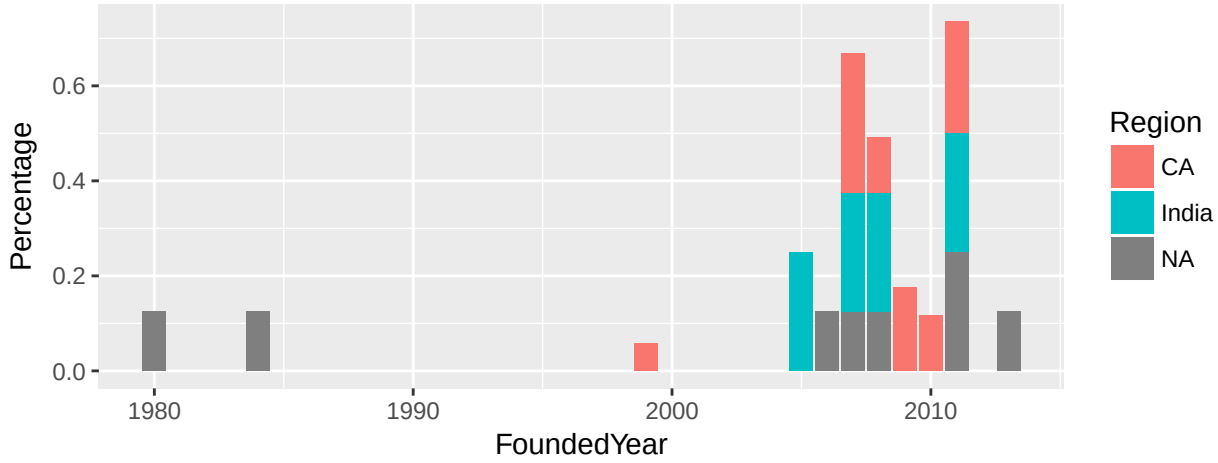

Figure 2: Distribution of Company ages

Linear growth in company value

All companies in the portfolio except one show some increase in their worth since the initial investment. A safe assumption, therefore, is that a correlation exists between their past and current worth. Below, the companies' current worth are regressed on their initial worth at the time of initial investment (Figure 5). Both numbers are scaled down by 100 million dollars. The most recent four companies that received CapitalG funds in 2017 are excluded. As expected, there is a very strong linear relationship between the initial worth and the current value of the company ($R^2_{adj} = 0.986$). The residuals of this linear fit are plotted at the Figure attached at the end of this document.

```
##
## Call:
## lm(formula = Current ~ Value, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.913  -5.801  -3.465   1.337  35.490
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.77625    2.68687    2.15   0.0454
## Value        1.00239    0.02714   36.93   <2e-16
##
## Residual standard error: 10.64 on 18 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.987,  Adjusted R-squared:  0.9863
## F-statistic:  1364 on 1 and 18 DF,  p-value: < 2.2e-16
```

The linear fit suggests that the company's current (future) valuation increases by a dollar with a dollar increase in its initial worth. This rate of change is mostly due to outlier companies. Figure 3 shows that the apparent outliers with extremely high current worth are Snap Inc., Airbnb, and Stripe. These companies receieved CapitalG Funding in 2016, and they have very high valuation figures but did not show significant growth over a year. Excluding these possible outlier observations significantly reduces the explanatory power due to the smaller sample and increased variance.

```
##
## Call:
## lm(formula = Current ~ Value, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.795  -8.125  -1.063   3.006  29.412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.6246     6.6755  -0.094   0.9269
## Value         1.8750     0.7123   2.632   0.0207
##
## Residual standard error: 11.1 on 13 degrees of freedom
## Multiple R-squared:  0.3477, Adjusted R-squared:  0.2975
## F-statistic: 6.928 on 1 and 13 DF,  p-value: 0.0207
```

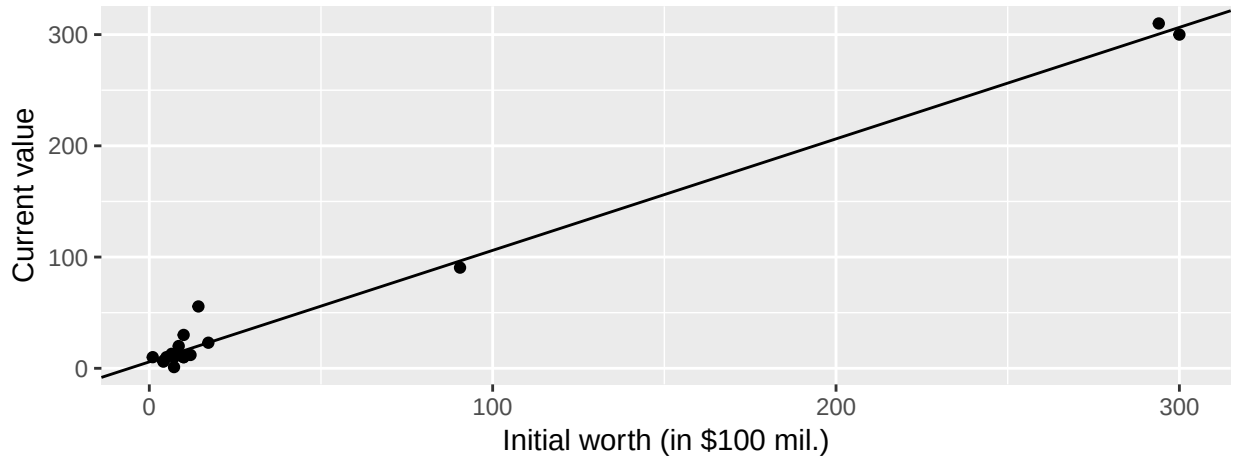

Figure 3: Current value on initial worth

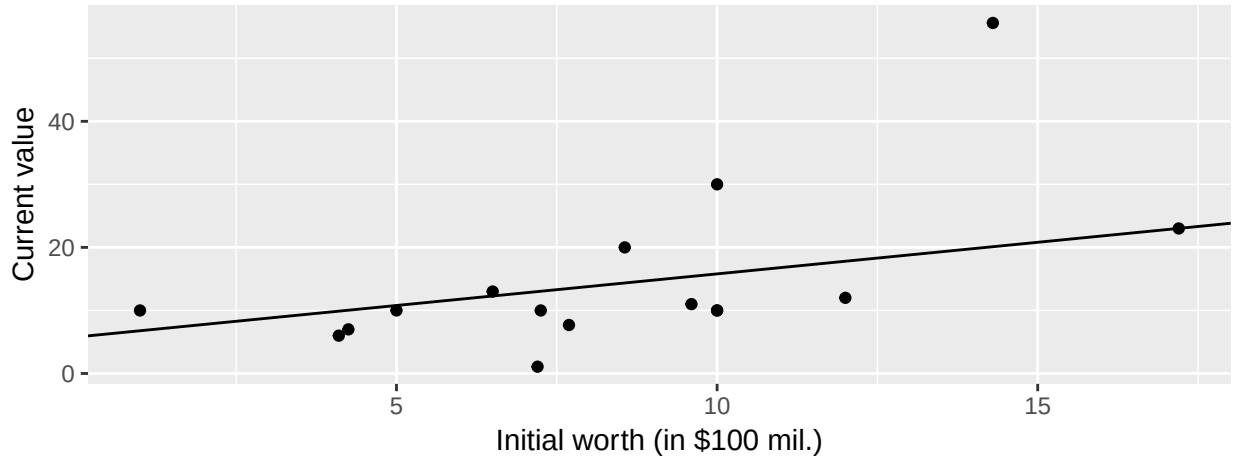

Figure 4: Current value on initial worth (zoomed in)

Relationship between Funding raised and Growth

Another possible relationship can be that between the amount of funding raised (from CapitalG and other lead investors) and the company's growth. Here we use the notion of company's growth as the its average worth increased since CapitalG investment. The most recent four companies that received CapitalG funds in 2017 are excluded.

No apparent pattern is discernible in the plot (Figure 5 below). Surveymonkey and Lyft received more than 4 billion dollars of funding and are clear outliers. Lending Club is an outlier company in terms of its high average growth compared with the funding it received.

```
## # A tibble: 4 x 3
##   Company       fund growth
##   <fct>        <dbl>  <dbl>
## 1 Surveymonkey  4.44  0.817
## 2 Lending Club  1.25  6.88
## 3 Airbnb        5.55  2.00
## 4 Lyft         10.0   0
```

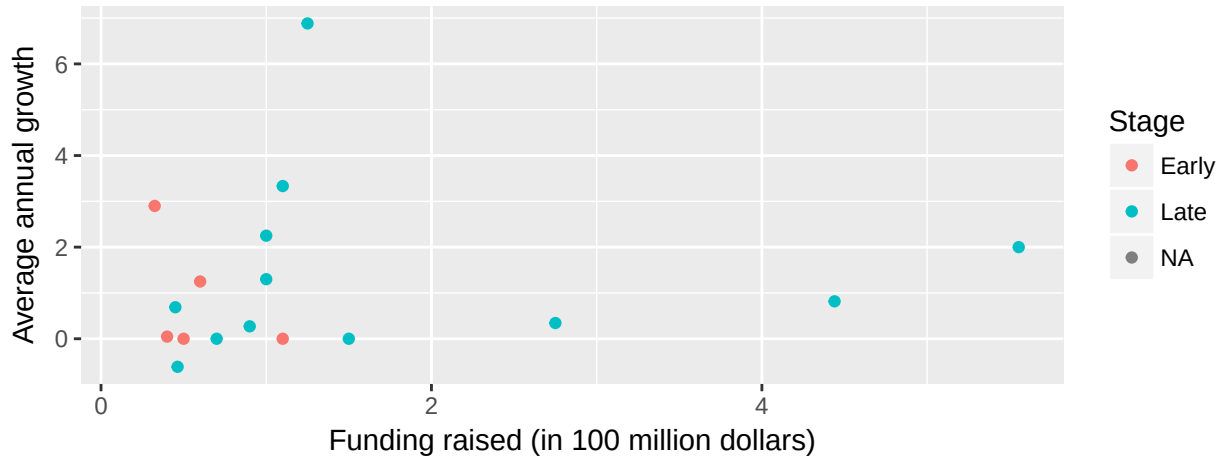

Figure 5: Growth plotted against Funding raised

Intuitively, such a plot with such few samples resist simplistic linear fit. As expected, any linear regression model has extremely low explanatory power ($R^2 \approx 0.02$). This result requires further examination, since it impplies that there is no clear relationship between the funding a company raised (which is closely linked to its valuation at the time) and its growth. A more detailed analysis using a more expansive data set will be able to address this issue more accurately and will help impprove future Investment Team decisions.

A Figure that plots the two linear fits (with and without outlier companies that received more than 2 mil. dollars) is attached at the end of this document.

Further questions

This analysis is obviously limited by constraints in time and resources. Addressing them will ultimately help us construct a model that can more accurately identify the investment patterns of CapitalG and predicts future investments of CapitalG for a certain company. These issues include:

i. Improving the data set, both in precision and in the variables included
ii. Trend analysis of company revenues over years
iii. More comprehensive modelling of company growth
iv. Identifying possible patterns/trends in CapitalG investment decision given the company's size, value, and revenues

---

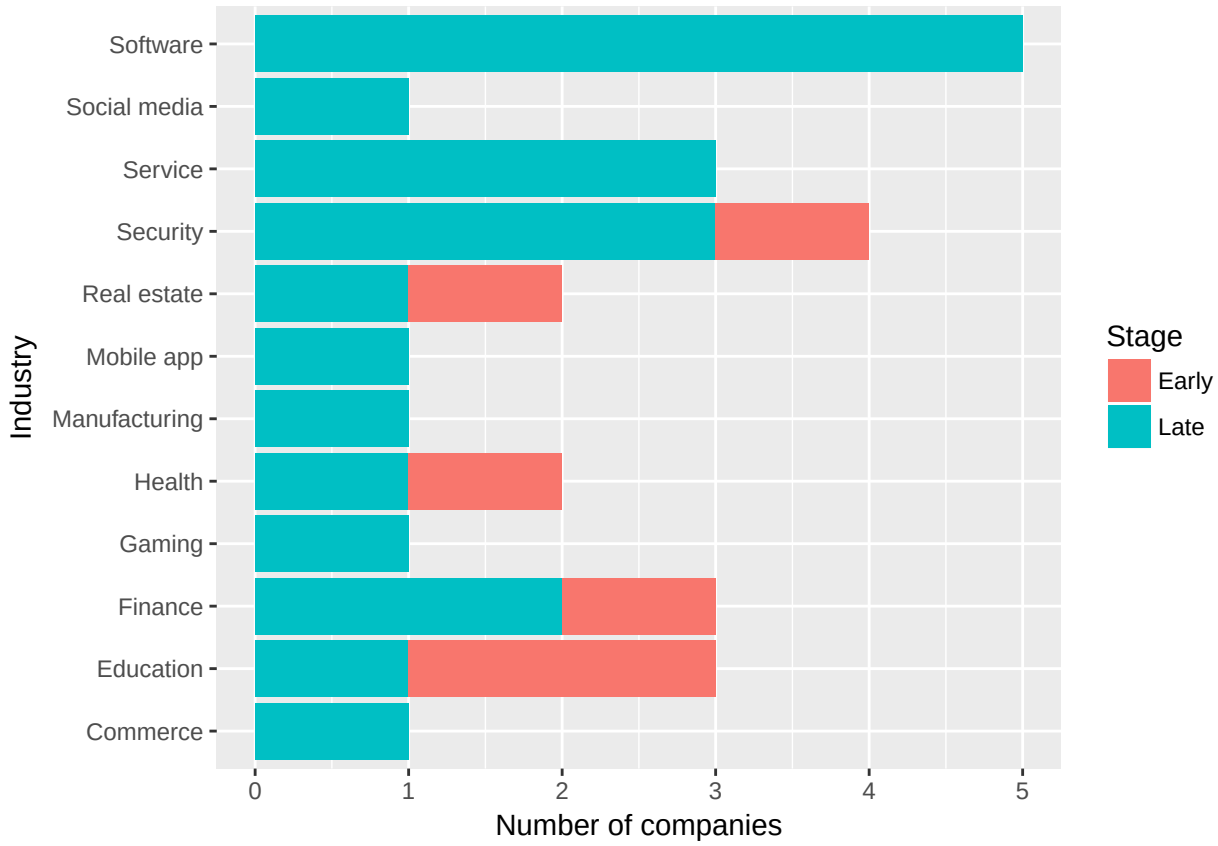

Figure 6: Industry distribution

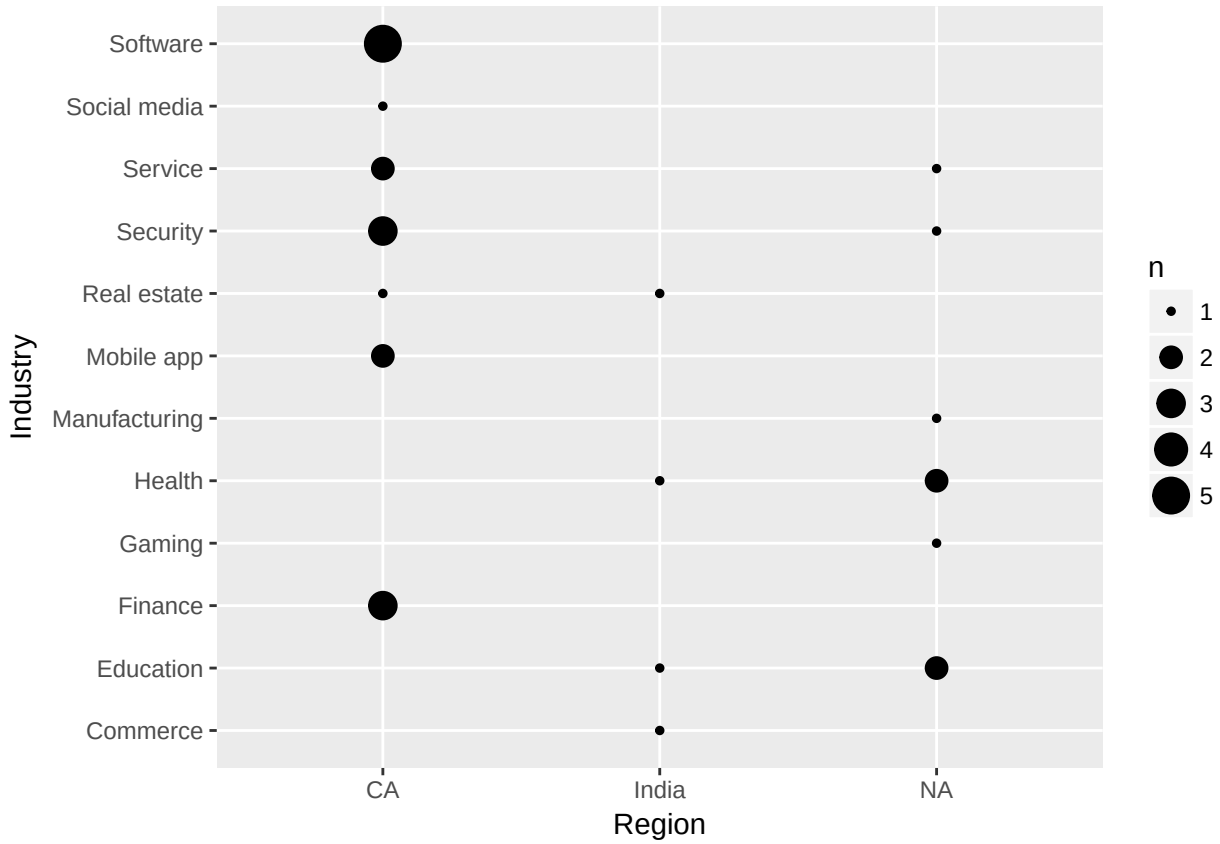

Figure 7: Industry distribution across regions

```
## Warning: Removed 8 rows containing missing values (geom_point).
```

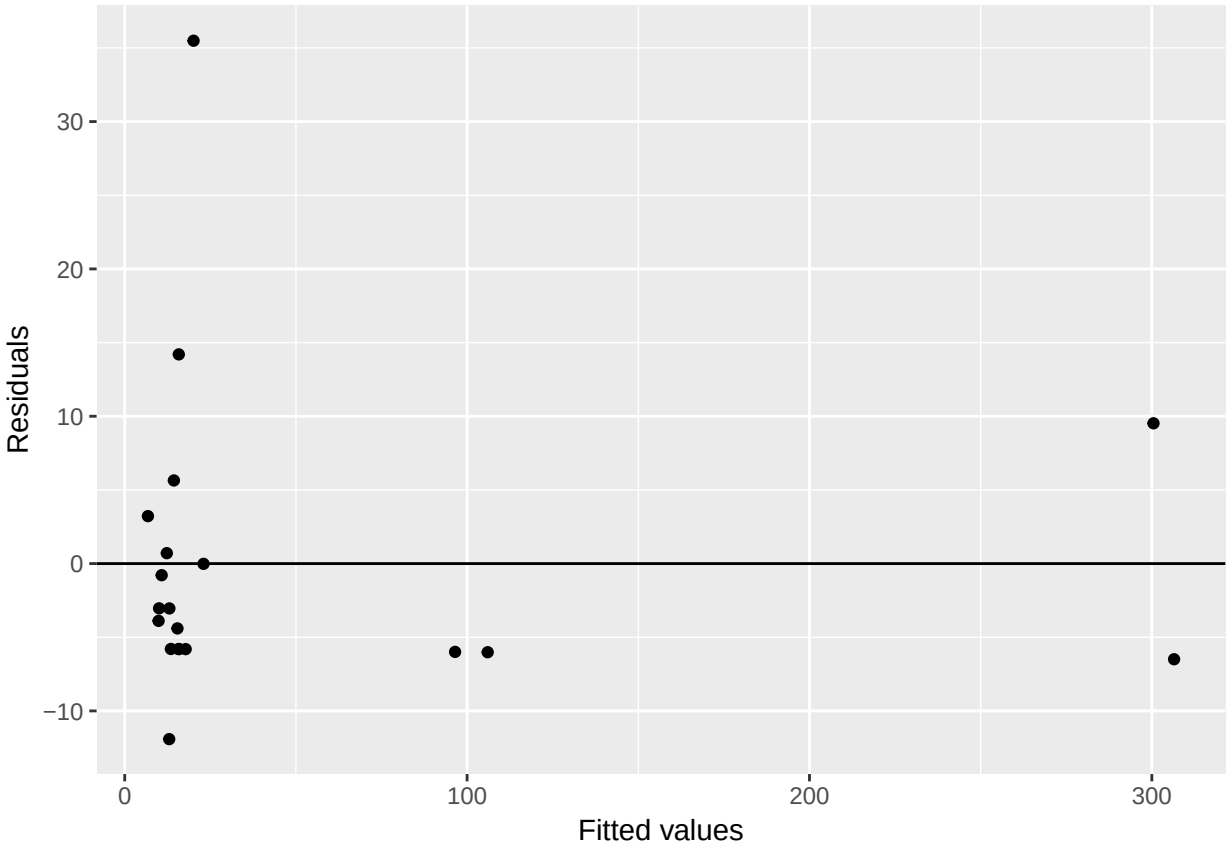

Figure 8: Residual plot

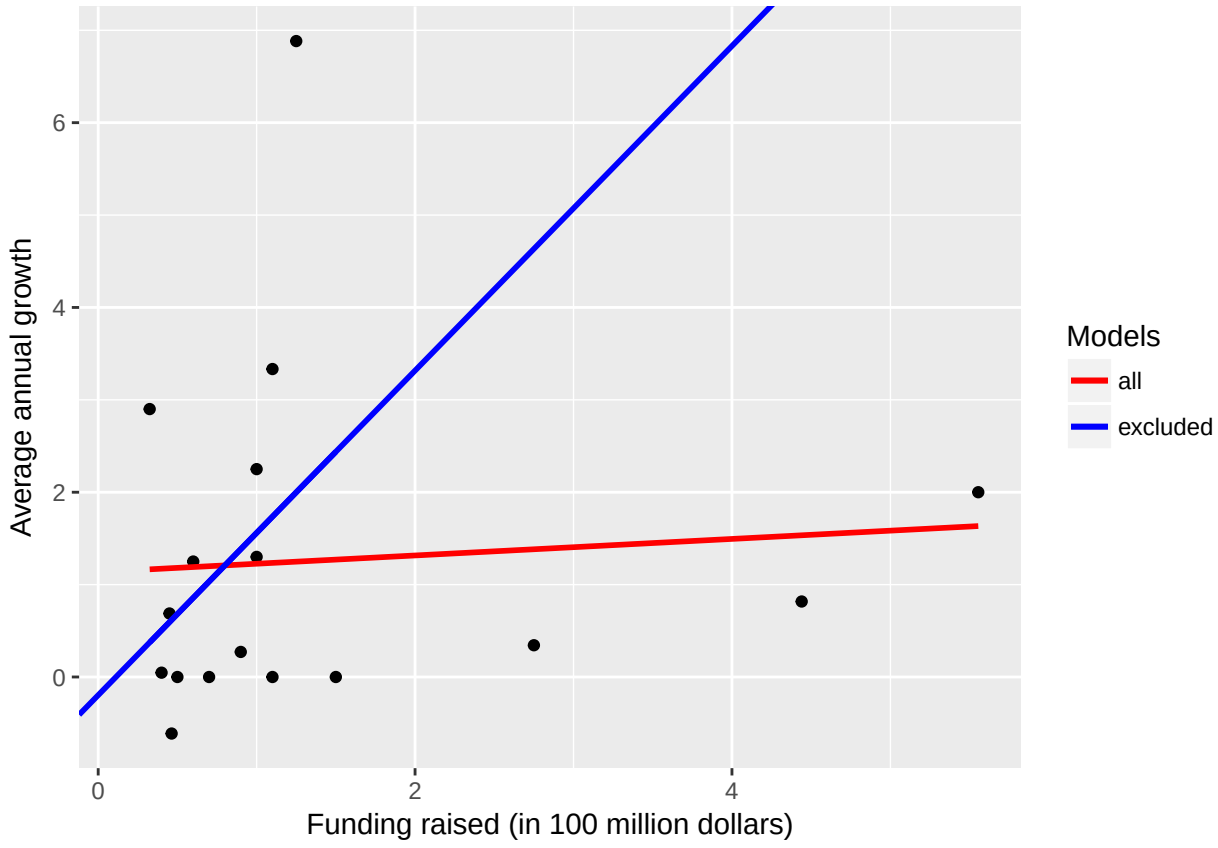

Figure 9: Linear fit to Growth on Funding