R&D expense and stock price

Kaiwei Xiao

Fall 2022

Abstract

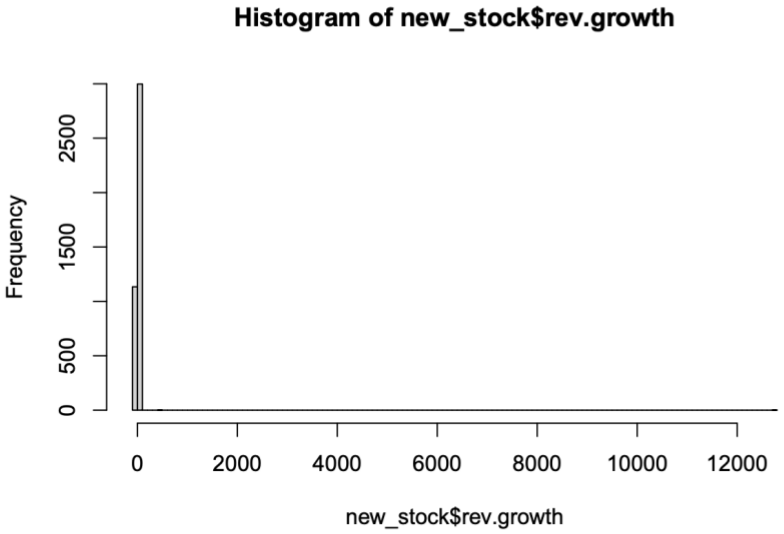
The stock market is quite complicated but still attracts lots of people to try to gain profit there. Some people and companies are trying to predict and find out what factors lead to the change in stock prices. In this project, there is a mixed model that tries to find out how R&D affects stock investment gains differently in 10+ sectors. The dataset comes from Kaggle and it contains the 2018 year US stocks with hundreds of predictors. However, we only focus on the R&D expense of these companies and include the revenue growth rate here for a simpler model building.

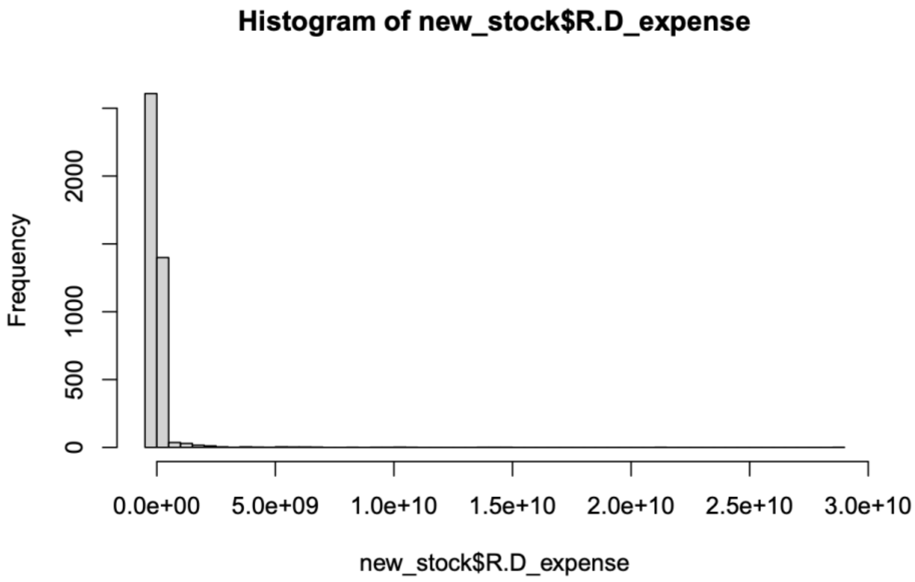
Introduction

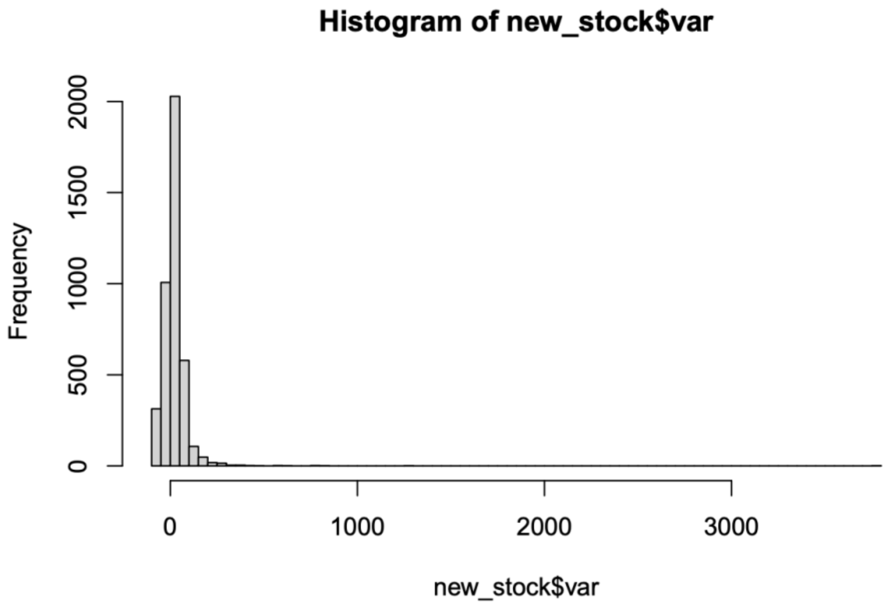
In my opinion, in the long term, how the company spent its money on research and development should have lots of effects on companies' earnings. And I come up with an assumption that different sectors can have different reactions to R&D expenses. Because for some sectors, the companies need to spend a lot on R&D because the techs are updated fast, and they need to compete with rivals heavily based on knowledge and techs. I also plan to include revenue growth in my model because companies need revenues to invest in R&D. At first, an intuitive guess is that Technology companies should be affected by R&D expenses much more than others.

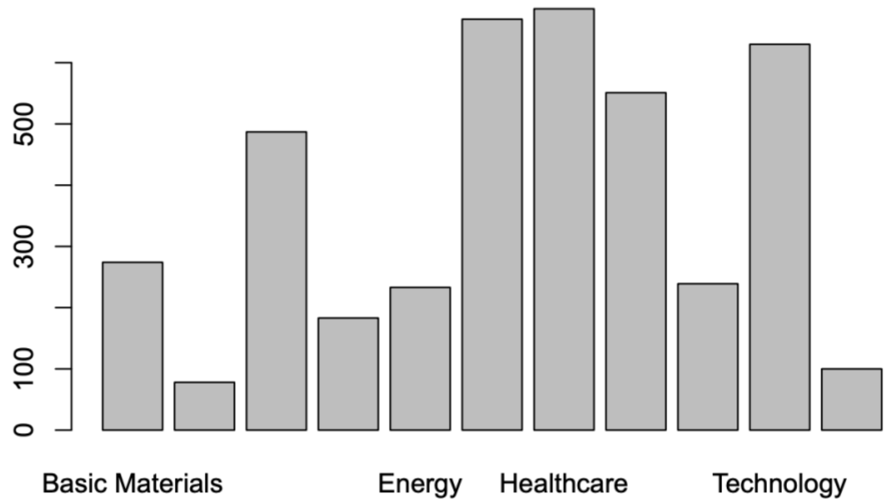
Methods

The data is pulled from Kaggle. Link: https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018?resource=download. I only used the 2018 data because I tried to avoid the panel dataset which I am not familiar with. The dataset contains thousands of companies with some NA values. Since the dataset is big enough and there are sufficient samples in each sector, I simply delete the NA rows. I first check the distribution of the variables including sectors, var(the gains or loss of stock price in 2018), the R&D expense, and revenue growth.



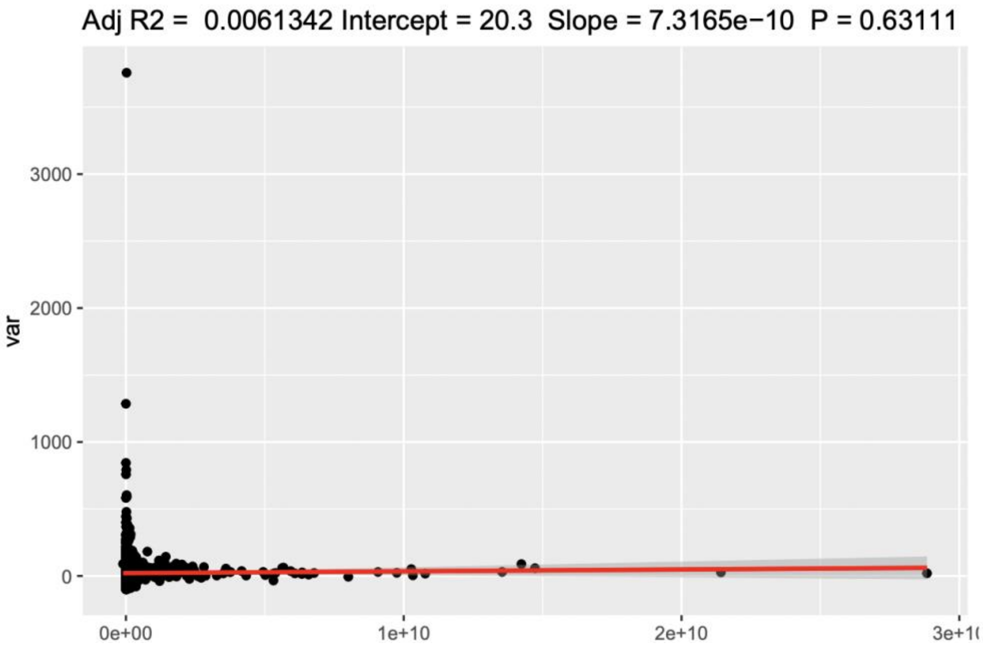

Histogram of new_stock$R.D_expense
Frequency
0
500
1000
2000
0.0e+00
5.0e+09
1.0e+10
1.5e+10
2.0e+10
2.5e+10
3.0e+10
new_stock$R.D_expense


Histogram of new_stock$var
Frequency
0
500
1000
1500
2000
0
1000
2000
3000
new_stock$var

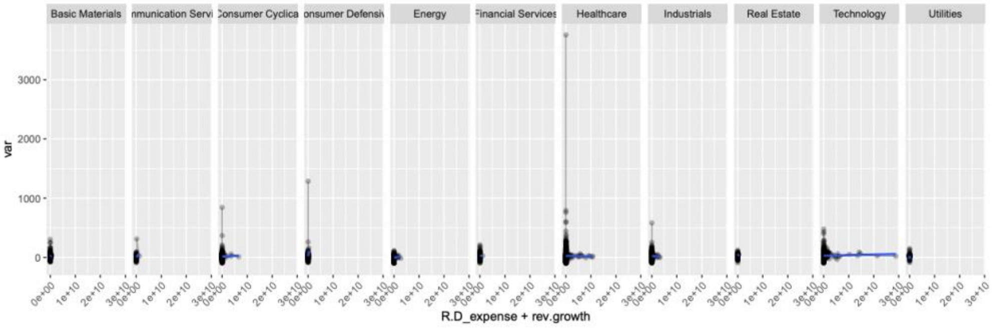


It is hard to say they are approximately normal distribution which may affect the robustness of the upcoming model.

Then I fit a linear regression model for exploration purposes. The model use var as the dependent variable and the rest as independent variables.

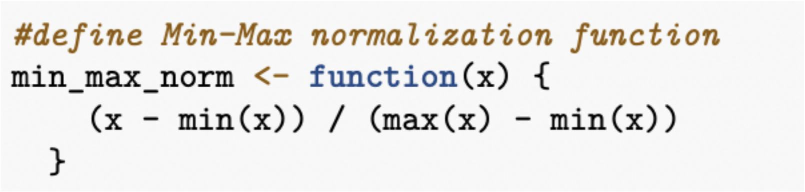


Adj R square is very small here which is reasonable because we omit lots of variables in the raw datasets. We can also see that we have to pay attention to the data scales since the dots are clustered.

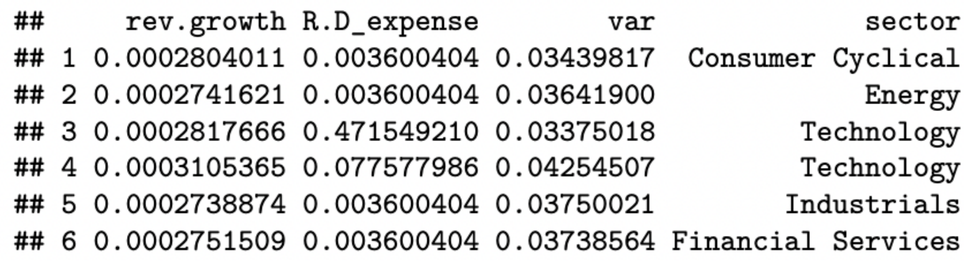


different sectors dots figure

Then I come up with a mixed model to test my assumption. It is a random slope and random intercept mixed model. I used the sector as a group. I normalize the data.

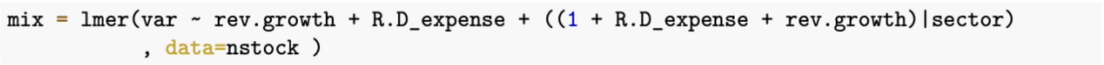


```
#define Min-Max normalization function
min_max_norm <- function(x) {
    (x - min(x)) / (max(x) - min(x))
  }
```



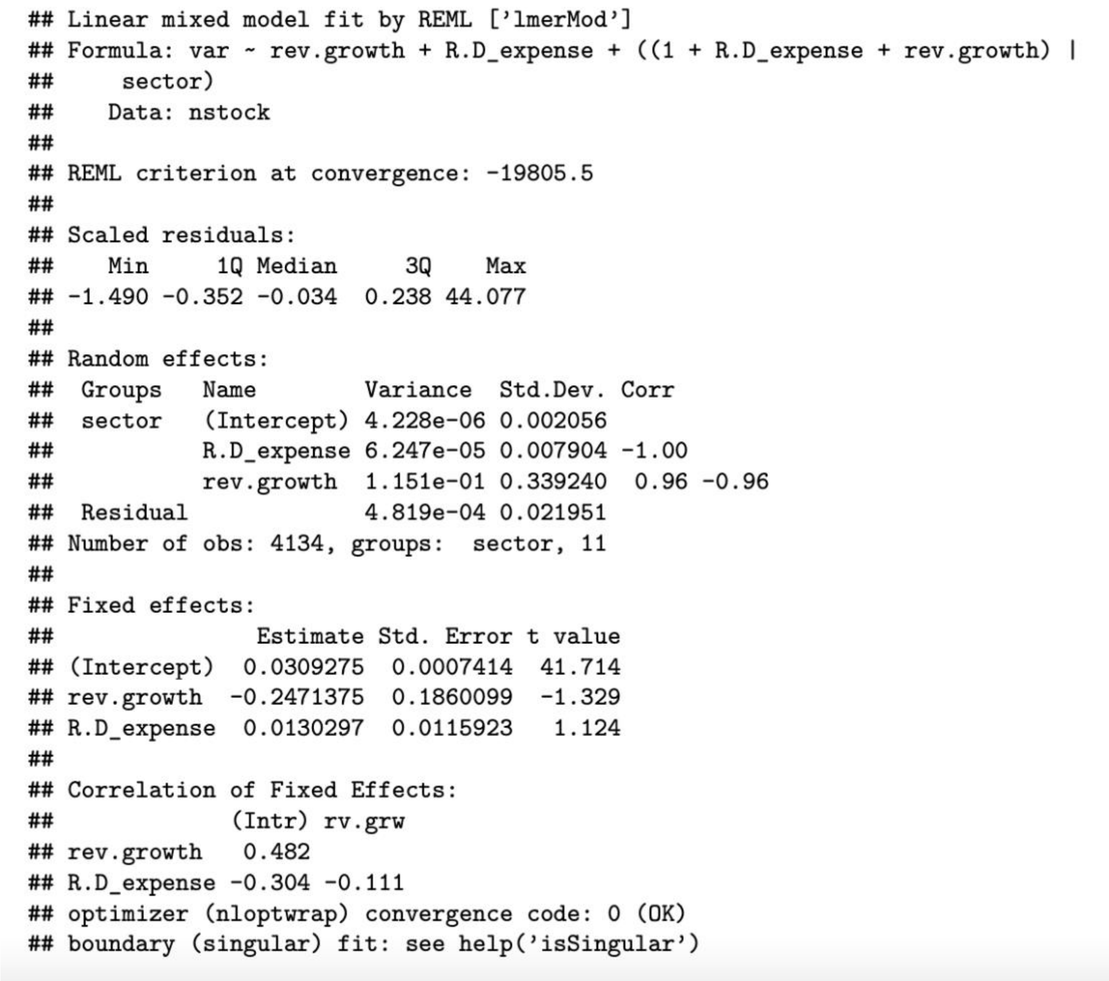
```
##      rev.growth R.D_expense        var                sector
## 1 0.0002804011 0.003600404 0.03439817  Consumer Cyclical
## 2 0.0002741621 0.003600404 0.03641900             Energy
## 3 0.0002817666 0.471549210 0.03375018         Technology
## 4 0.0003105365 0.077577986 0.04254507         Technology
## 5 0.0002738874 0.003600404 0.03750021        Industrials
## 6 0.0002751509 0.003600404 0.03738564 Financial Services
```

Model building

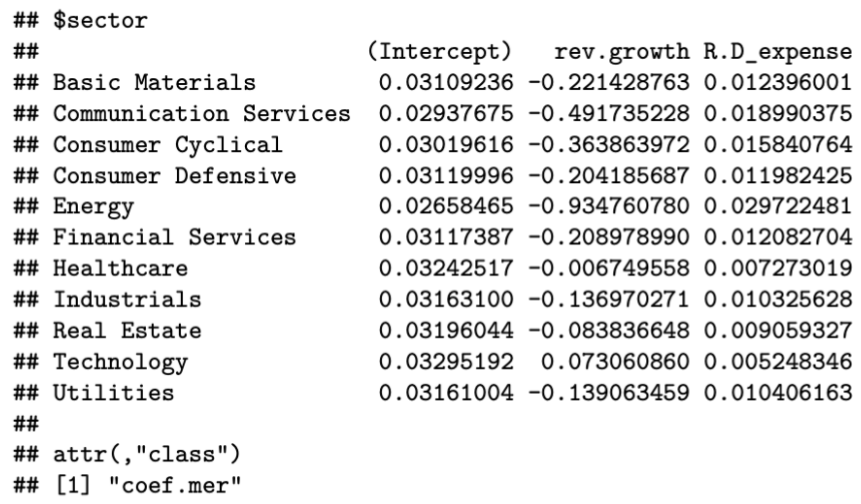


```
mix = lmer(var ~ rev.growth + R.D_expense + ((1 + R.D_expense + rev.growth)|sector)
           , data=nstock )
```

Results



```
## Linear mixed model fit by REML ['lmerMod']
## Formula: var ~ rev.growth + R.D_expense + ((1 + R.D_expense + rev.growth) |
##     sector)
##    Data: nstock
##
## REML criterion at convergence: -19805.5
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -1.490 -0.352 -0.034  0.238 44.077
##
## Random effects:
##  Groups   Name        Variance  Std.Dev. Corr
##  sector   (Intercept) 4.228e-06 0.002056
##           R.D_expense 6.247e-05 0.007904 -1.00
##           rev.growth  1.151e-01 0.339240  0.96 -0.96
##  Residual             4.819e-04 0.021951
## Number of obs: 4134, groups:  sector, 11
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  0.0309275  0.0007414  41.714
## rev.growth  -0.2471375  0.1860099  -1.329
## R.D_expense  0.0130297  0.0115923   1.124
##
## Correlation of Fixed Effects:
##             (Intr) rv.grw
## rev.growth   0.482
## R.D_expense -0.304 -0.111
## optimizer (nloptwrap) convergence code: 0 (OK)
## boundary (singular) fit: see help('isSingular')
```



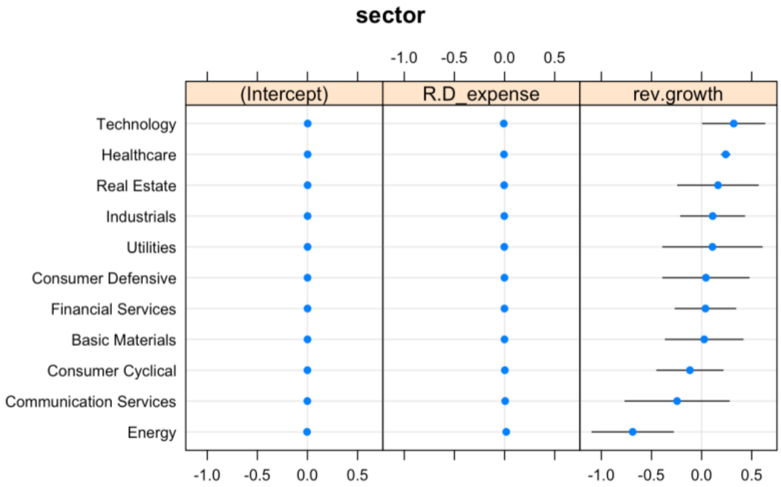
```
## $sector
##                        (Intercept)   rev.growth R.D_expense
## Basic Materials         0.03109236 -0.221428763 0.012396001
## Communication Services  0.02937675 -0.491735228 0.018990375
## Consumer Cyclical       0.03019616 -0.363863972 0.015840764
## Consumer Defensive      0.03119996 -0.204185687 0.011982425
## Energy                  0.02658465 -0.934760780 0.029722481
## Financial Services      0.03117387 -0.208978990 0.012082704
## Healthcare              0.03242517 -0.006749558 0.007273019
## Industrials             0.03163100 -0.136970271 0.010325628
## Real Estate             0.03196044 -0.083836648 0.009059327
## Technology              0.03295192  0.073060860 0.005248346
## Utilities               0.03161004 -0.139063459 0.010406163
##
## attr(,"class")
## [1] "coef.mer"
```

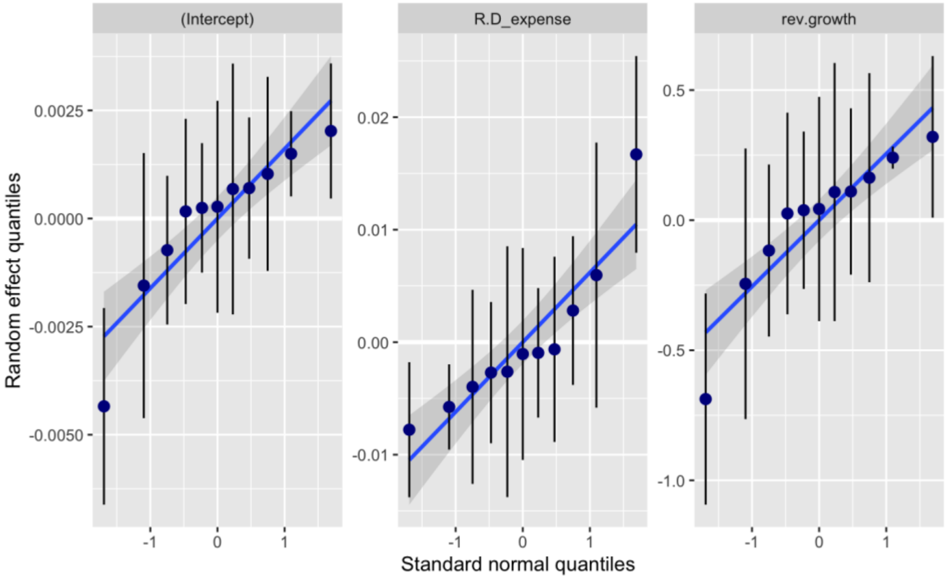
It is hard to fit this mixed model unless we did the normalization. Because the raw data have very different scales. And some companies did not invest in R&D. Also, the

REML method is used here. If we use the maximum likelihood method here, we will get different results. REML method is chosen because usually it provides a more precise result.

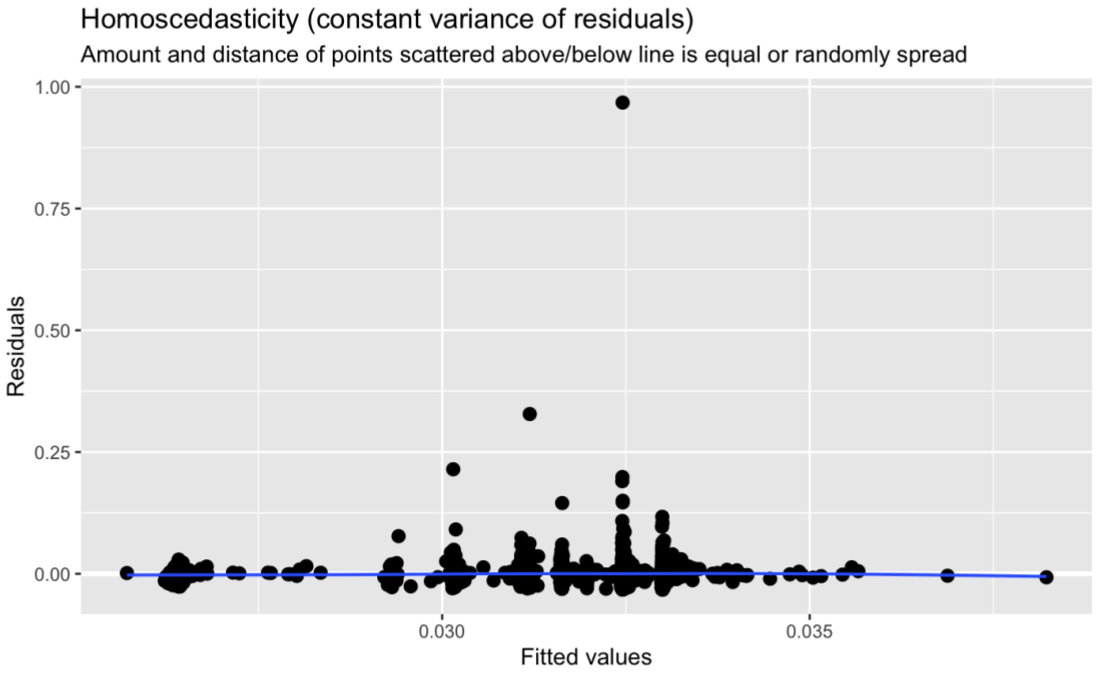
Since we rescale the data, so what we can say here is only the relative relationship. It is surprising that the coefficient of revenue growth is minus. Usually, investors tend to invest in higher revenue growth companies. And the positive coefficient of R&D expenses makes sense. When companies invest more in R&D, the market receives news and investors probably buy their stocks which leads to a higher stock price.

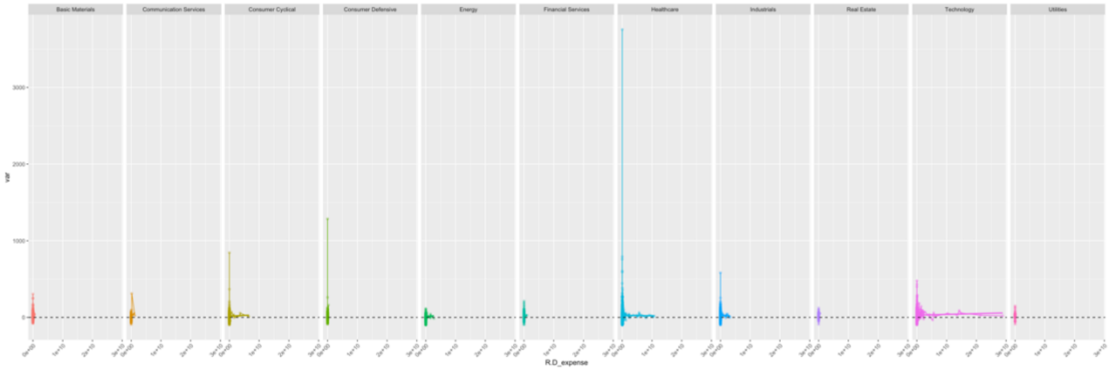
For the random effects, the Technology companies surprisingly give us a 0.0052 R&D expense coefficient which is relatively small compared to other sectors. A simple guess is these technology companies need time to utilize new technologies and make new products. So there might be a lag for their stock price to go up. But still, we can see different sectors do have different sensitivity to R&D expense and revenue growth.





The random effect QQ plots are ok here.





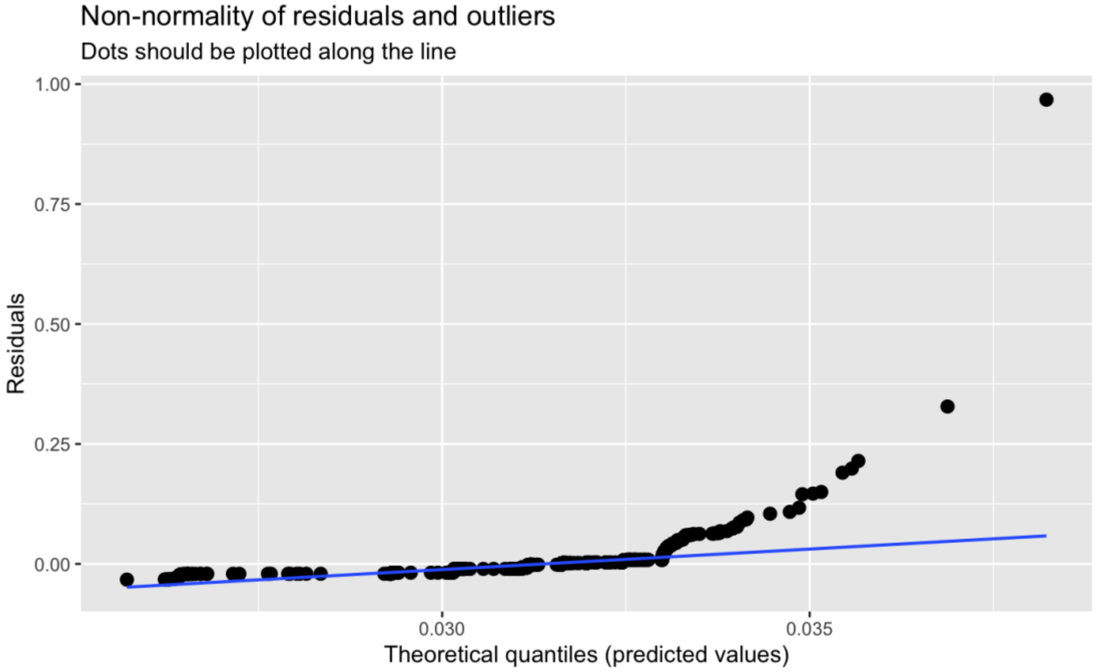
The residual plots are not perfect. I checked the raw data; this is probably due to some huge rise in the stock price of healthcare companies. But overall, the residuals look ok.

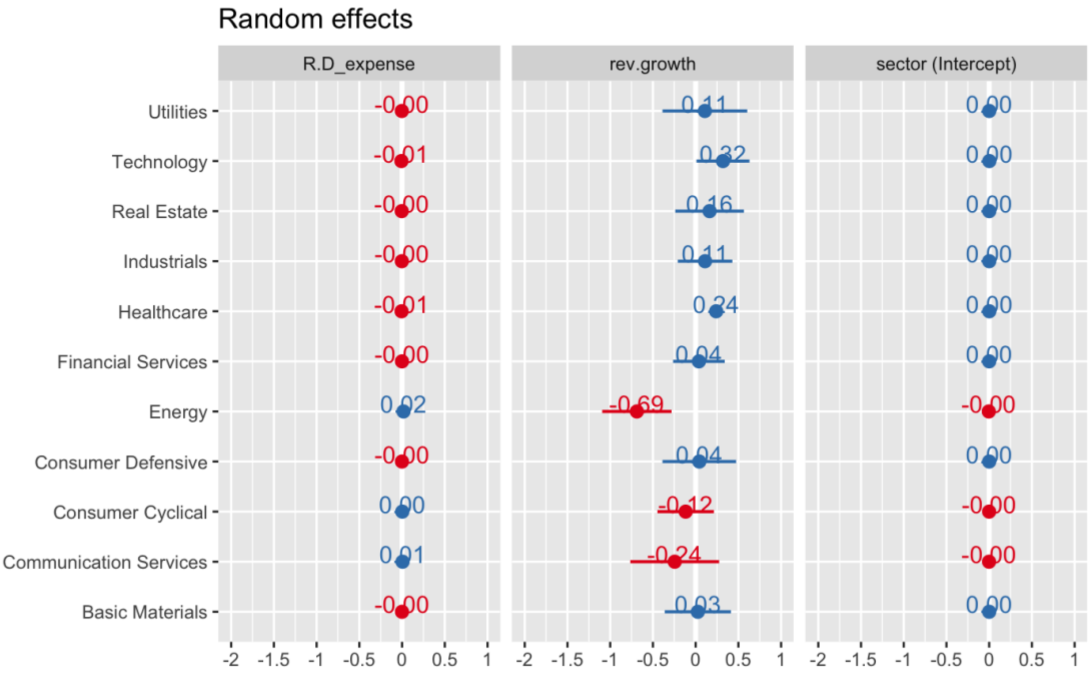
Discussion

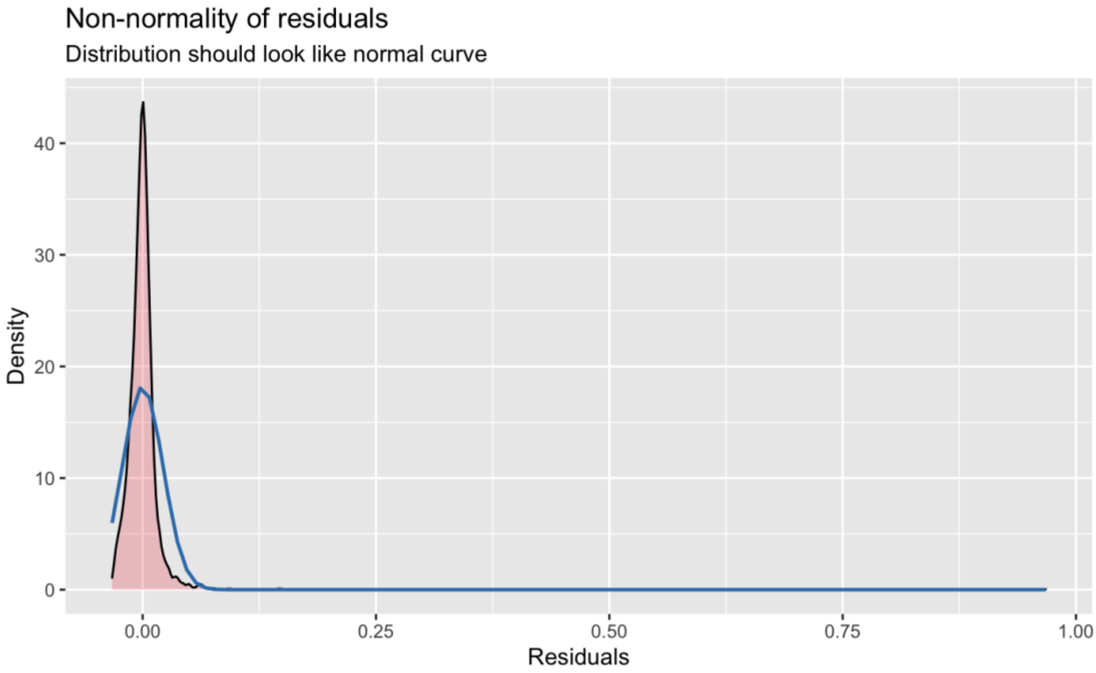
Further improvement can be done. For example, a better rescale formula might help us find the coefficients that are easy to interpret in the real world. We can include the net profit variables in the model to see how they interact with other variables. Overall the stock market is really complicated and it is hard to make a perfect model that predicts the trend for us. But at least we can say it is always a good choice for the company to invest in R&D and stock market investors should look for companies that tend to spend more on their R&D departments.

Appendix

More model checks







References

Kaggle dataset link: https://www.kaggle.com/datasets/cnic92/200-financial-indicators-of-us-stocks-20142018?resource=download