Jake Gadaleta | Regression

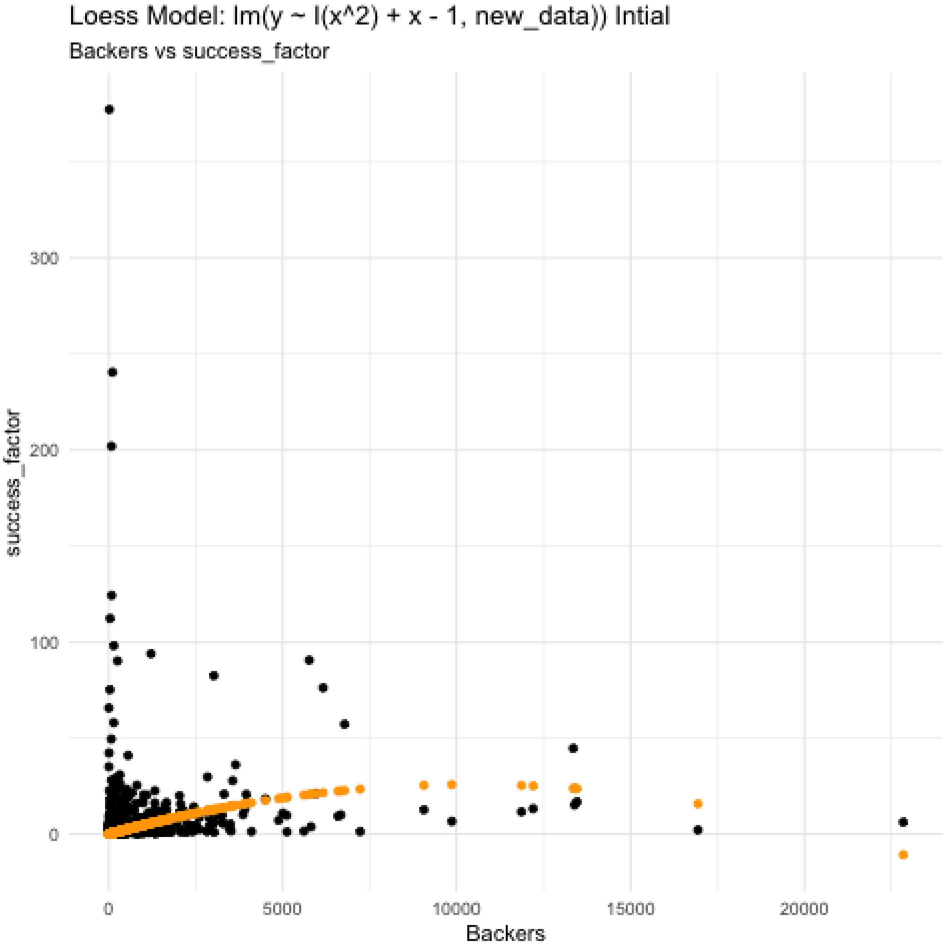
Before we get into things I want to take the time to specify that for these regressions I have sampled 10,000 rows from the original dataset (excluding outliers). This change was made to boost compile times and also make sure that the graphs were actually readable. Attempting to create these graphs using all 300,000 rows took ~10 hours (and then there was still debugging to be done). As always the source code, and higher definition images are available on github

For both of these regressions I have chosen to generate Success Factor as I did in the last section. I also took time to look over the data once again, both of these algorithms use Backers as their other factor. Backers was chosen because it is the only numerical variable that is not used in the construction of Success Factor.

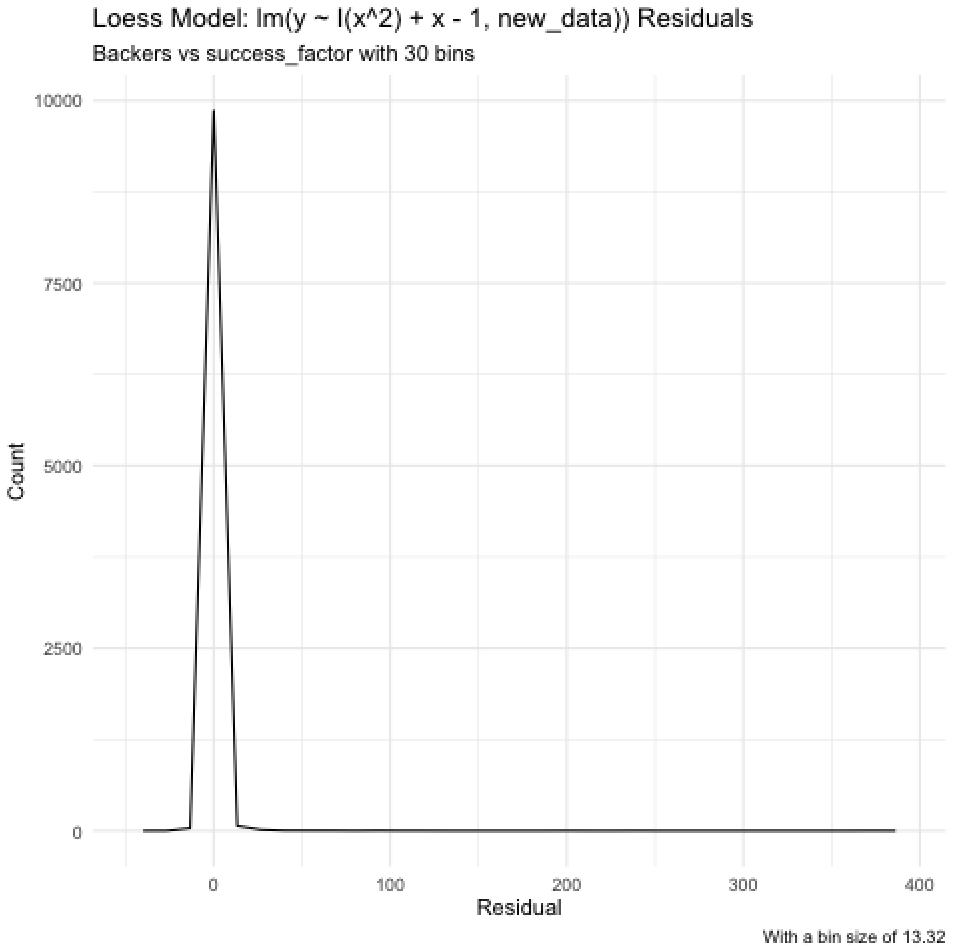
It should also be noted that the vast majority of KickStarter Projects actually do fail which has made this project way less fun than I really Expected it to be.

Linear Model

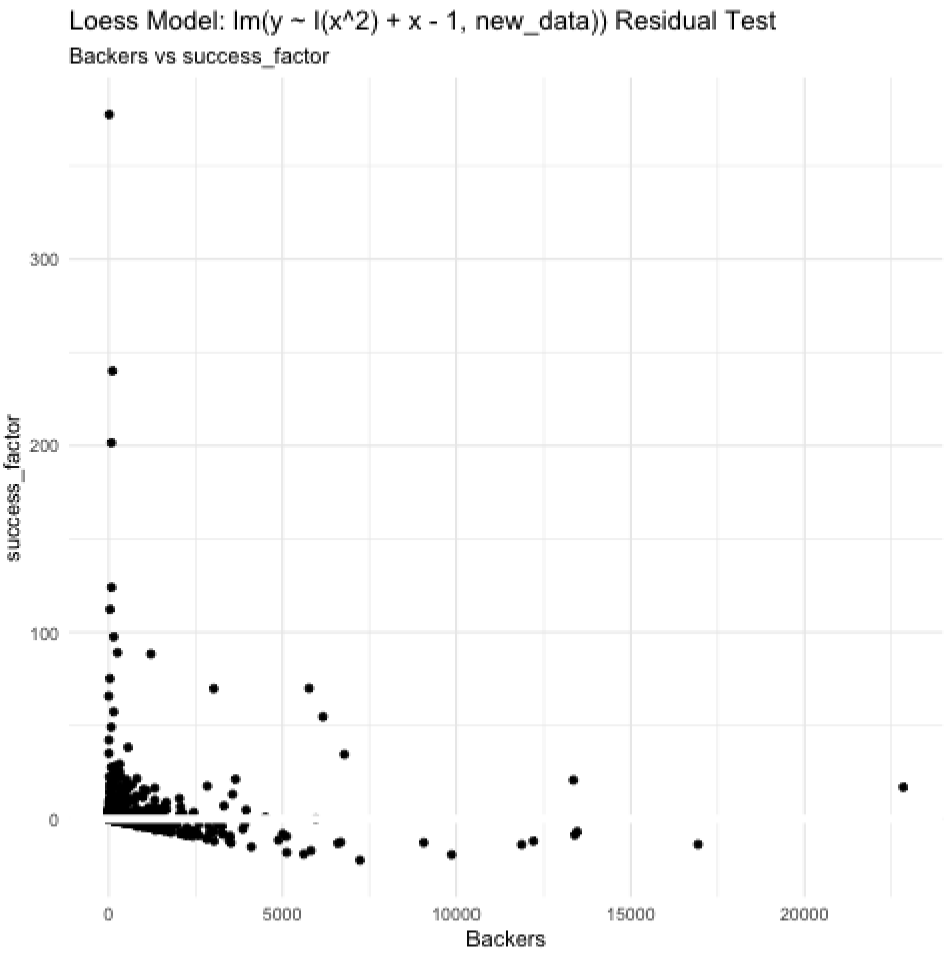
Success Factor ~ I(x^2) + x-1



here on our initial graph we do seem to have some strong predictability power in the beginning but quickly lose that as numbers increase. Please note that scale of model a Success Factor of 100 means that the project made 100 times over the amount of money that they asked. This means that not just this model is bad at expecting how much capital certain projects can bring in.



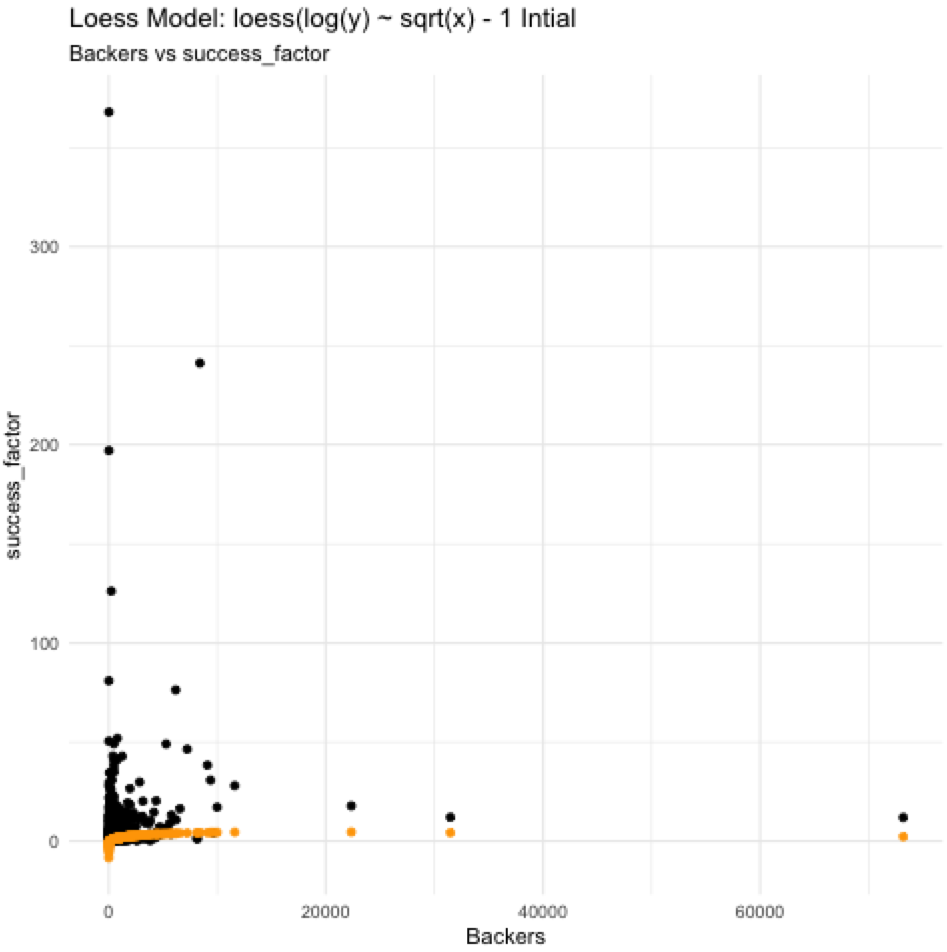
Now see our residual chart and again we see high success at +/- 13 but quickly we lose the ability ability to predict. I played with the bins here a lot and short of making it nearly 500 bins it was impossible to generate a graph that displayed decent data.



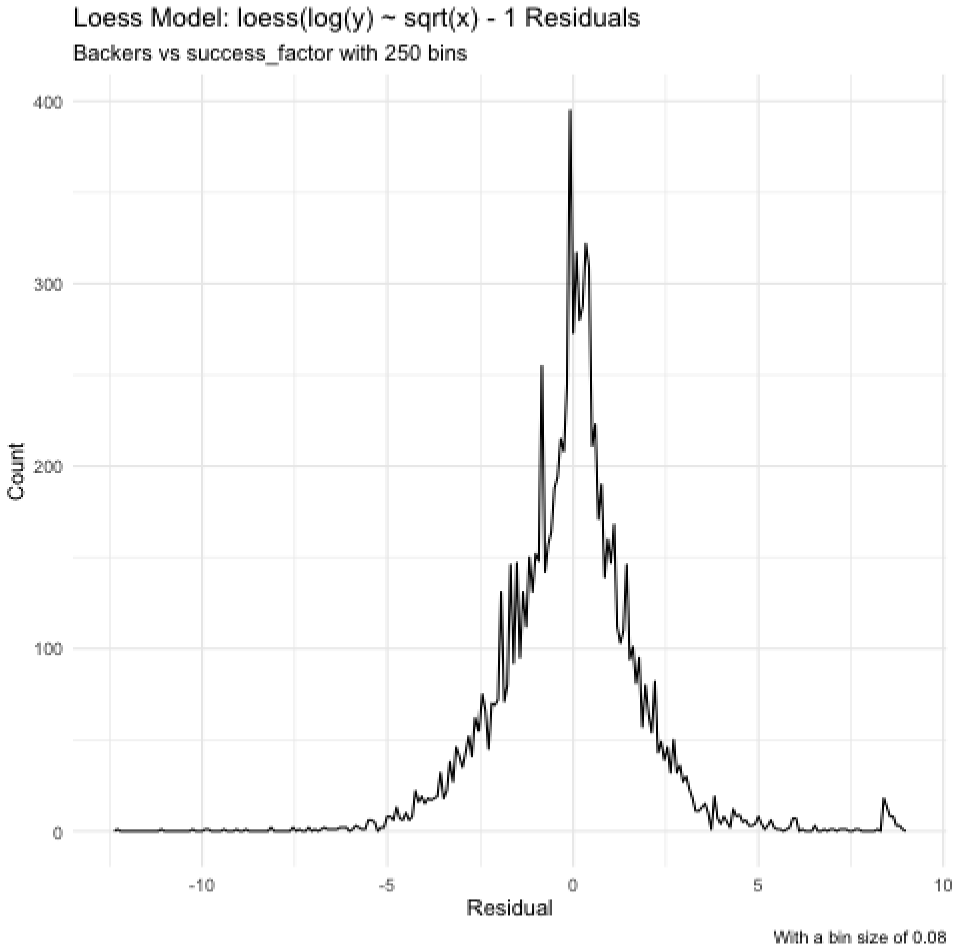
This isn't the best that it possibly could be interpreted as the spread isn't quite as good as I'd like it to be but it isn't linear so thats nice.

Loess Model

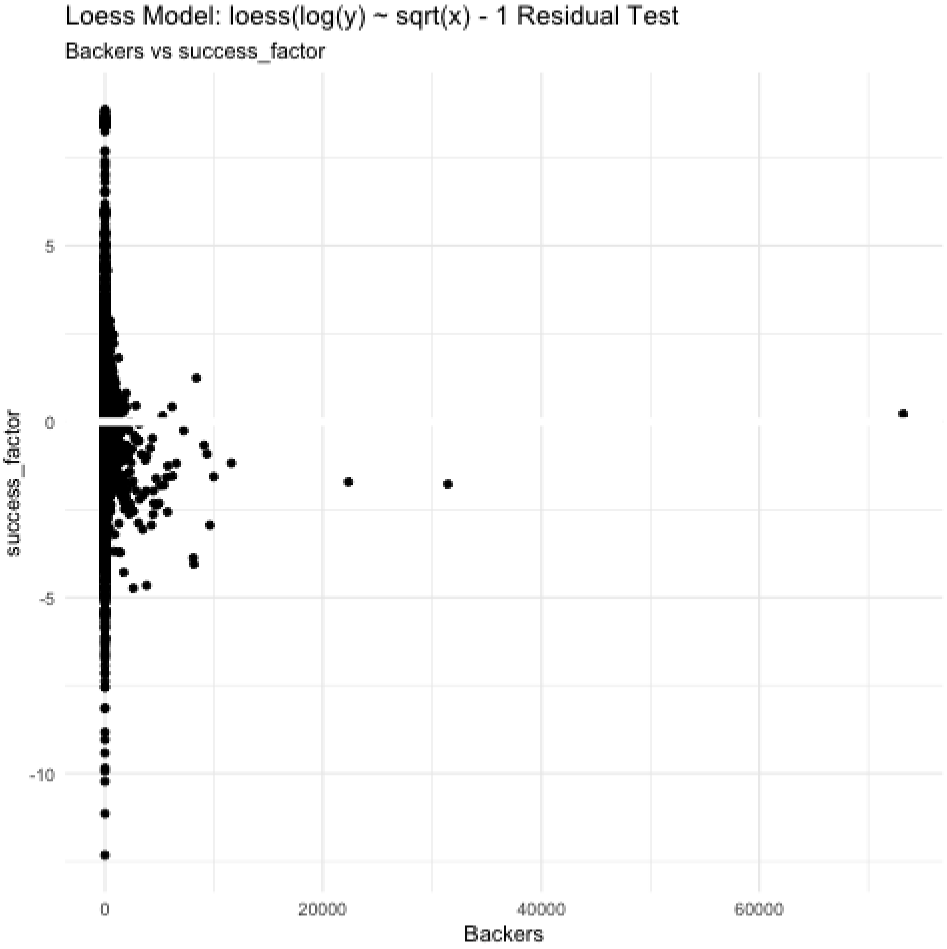
log(Success Factor) ~ sqrt(Backers) - 1



Yet again we are messed up by those high numbers but for a brief moment we seem at least a little accurate.



For this one I wanted to make sure that the residual plot showed some form of information so I made the bins very small (.08) from here we can also conclude again at small points its great but as we grow we lose a lot of that predictability



It sure ain't linear.

Conclusion

In conclusion throughout this entire project and across 3 separate class this data set fought me at every step of the way. If I had more time / taking less class I think I would've loved to work with some of the date information (launched, deadline) but most of the code that I had from the class didn't fit and I didn't have the time to learn it. I also generated a lot of graphs and even though some don't make sense and others don't show excellent information I have proven that I can script lots of tidyverse graphs.

I also want to take the time to thank you Prof, this has been a fun class and sometimes a release of stress for me during this crazy semester.