Simulating a Ghost-Writing Startup – A Sequel

by Jonathan Chen

Abstract: There is rarely a business endeavor that is profitable from the start. In this follow-up analysis, I detail using Arena to simulate the interactions between revenue, marketing costs, and the customer fanbase for a book series published on Amazon via ghostwriters. Using the takeaways from my previous paper and incorporating newly available marketing and sales data, my task is to determine the conditions under which our company will generate a profit. I first simulated how a standalone book behaves and tuned the model to mirror historical data. I then established some baseline assumptions to expand the model to handle multiple books in the same series. Ultimately, I find that profit is feasible. However, even under the relatively optimistic parameters used in the model, it requires a decent overhead investment of both time and money.

Background/Problem Description: After a previous analysis where I examined the financial feasibility of our company's pay structure and business operations, I realized that the nuances of buying a manuscript are overshadowed in importance by how it sells. I also concluded that marketing, whether done consistently or just to renew interest after novelty wears off, was an indispensable investment that must be made no matter how bullish I simulated customer traction to be. This meant that in this follow-up analysis I could simplify the model and focus solely on marketing costs and revenue. Factoring in these priori, my goal is to re-simulate our company's operations using less optimistic parameters. I will then need to determine under what circumstances the company will generate a profit with these new assumptions.

I will first construct a simple model with just one book and tune it to mirror the financials of an inaugural book in a series. This step serves as a reality check and will provide insight on profitability in the short term. After analysis I will then generalize to include multiple books and reexamine profitability.

Main Findings: One flaw of the previous model was that I underemphasized marketing as a cost component and only focused on book sales as a source of revenue. Revenue comes in two forms: royalty from book sales and royalty from pages read by premium Kindle subscription users. Instead of simulating the royalties of each one separately, one simplification was to simulate the combined sum.

One pivotal assumption was that the currency between marketing and sales is solely based on clicks generated from showing ads. This is a reasonable assumption considering all our company's marketing and sales are done digitally. This makes the number of clicks our ads get a function of expenditure and time. It can also be assumed that each click has some averaged out monetary value, or revenue per click (RPC).

To solve for the relationship between ad expenditure (ad spend) and clicks, I ran a simple linear regression between the two and included a monthly time component to identify any possible trend. The summary is below.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 371.2010    97.6706   3.801 0.005233 **
cost          0.9023     0.1396   6.465 0.000195 ***
months       -9.9608    13.3617  -0.745 0.477315
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 129.4 on 8 degrees of freedom
Multiple R-squared:  0.8711,    Adjusted R-squared:  0.8388
F-statistic: 27.02 on 2 and 8 DF,  p-value: 0.0002763
```

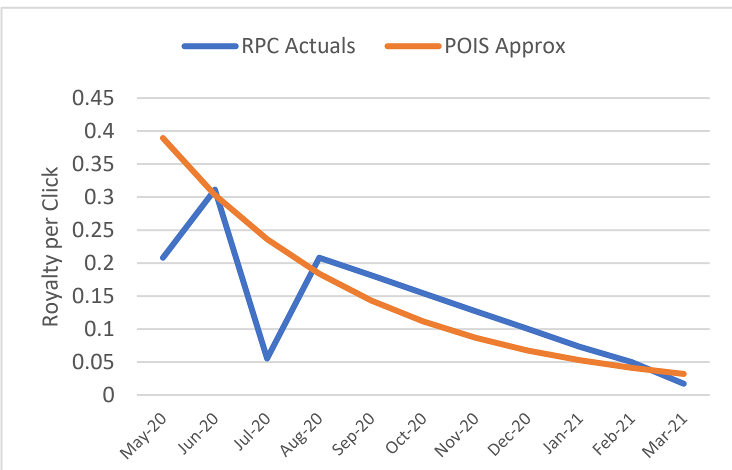
The intercept can be interpreted as the "novelty factor." Implying that you can get 371 extra clicks for an ad just because it is new.

The cost coefficient signifies that every dollar spent translates to .9 more clicks.

The coefficient for months implies that potential for clicks decreases by ten every month after initial release.

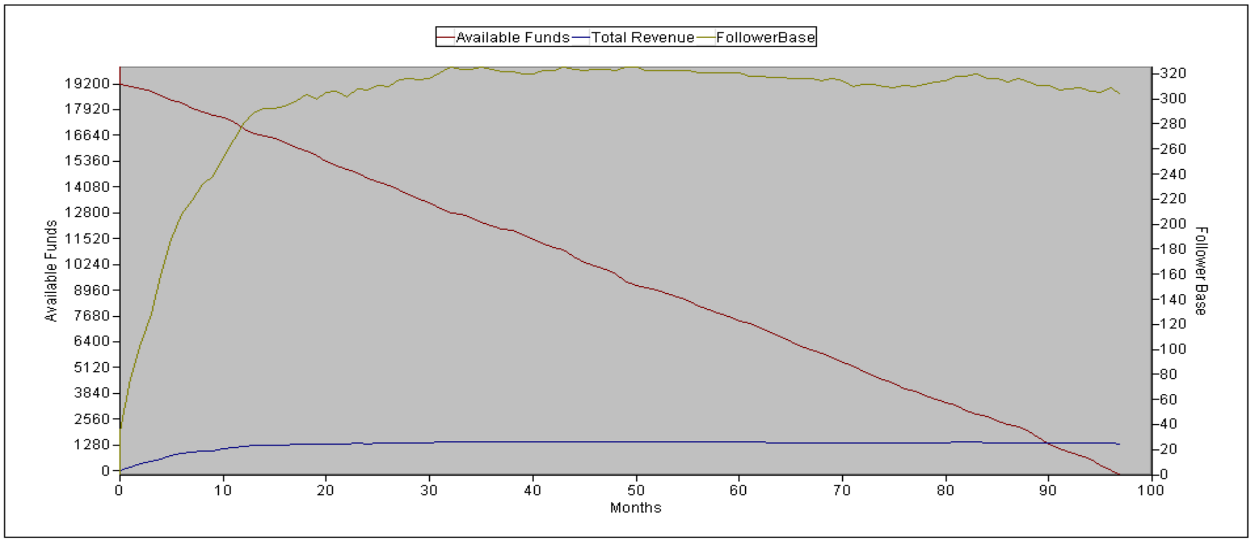
Even though only cost and the intercept are statistically significant to predicting clicks, I decided to include the monthly degeneration value as well. It makes sense that over time there will be a natural decline in ad effectiveness. Though in future analyses, I will need to validate whether this insignificance is due missing variables, or if monthly degeneration is truly a non-factor in predicting clicks.

Now that we have a way to simulate clicks over time, we need to look at RPC. The historical RPC is in blue while the function I used for approximation is in orange.



Similarly, to how I handled sales in the previous model, I approximated the decreasing behavior of RPC by using a Poisson distribution. Poisson distributions have interarrival rates that are exponential which take the form of: $\lambda e^{-\lambda t/\beta}$. Where lambda ($\lambda$) helps determine my starting RPC, 't' represents time that has passed, and beta (β) allows me to control how fast the function decreases. Specifically, I decided to make these parameters: $\lambda = .5$ & $\beta = 2$.

I now have all the pieces for an initial run. Like in the previous analysis, I initialized the amount of *Available Funds* to $20k. Having manuscript cost as a random variable did not have much impact on how the last simulation ran, so in this analysis I kept it as a flat value. Each manuscript will be bought for $2800 of which $700 will be paid upfront and $2100 will be paid as royalty payments overtime to the ghostwriter. To mirror the actual financial numbers, $140 in ad spend was used for pre-order exposure. Afterwards, ad spend was a normal random variable with: $\mu = 200$ and $\sigma = 50$. *Click Potential* follows the function: $371.2 + .9023 * AdSpend - 9.96 * MonthsPast$. Revenue was calculated by multiplying *Click Potential* with the current RPC, which follows the function: $.5e^{-5*\mathrm{MonthsPast}/2}$. Like in the previous analysis, 60% of the royalty will go towards paying off the remaining $2100 owed to the ghost writer, and 30% will be returned to the *Available Funds*. After the ghost writer has been fully paid off, 72% will go back into *Available Funds*. I ran this simulation until either 100 periods (months) passed or *Available Funds* ran out. Keeping track of total revenue earned, *Available Funds*, and *Follower Base* gives us this output. *Follower Base* is a metric meant to describe the number of people who have become loyal to the book series. It is not actively used in this first model, but I track it because it is pivotal in the next.

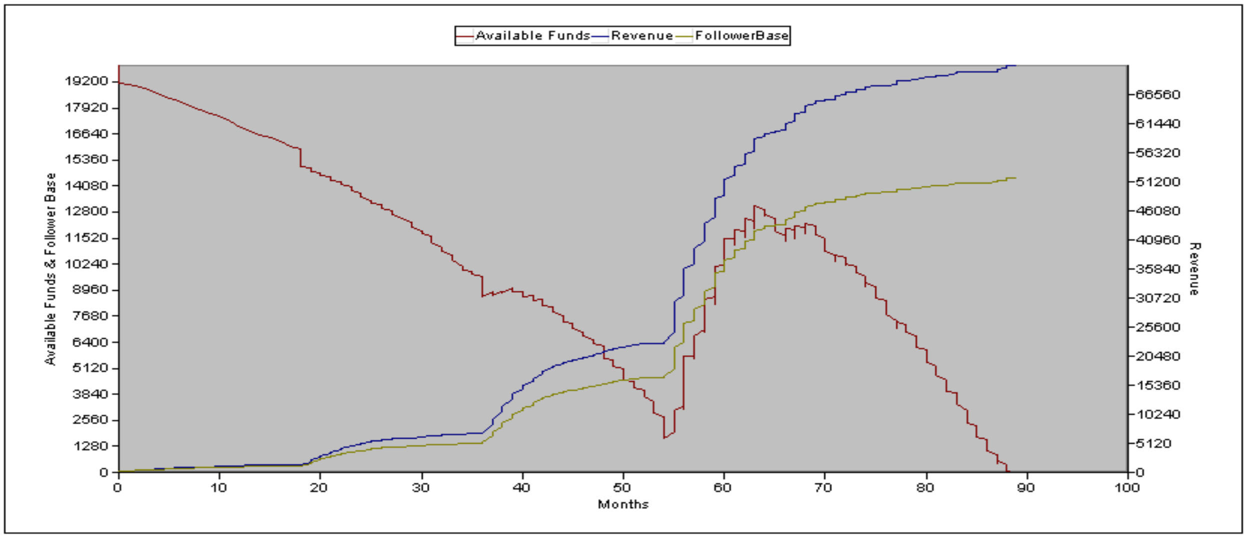


Forecasting over 80 periods from only 11 periods of real data will never yield accurate results. But the best we can do to confirm that we at least started on the right foot is by comparing the first 11 periods of this simulation to the 11 periods of actual data that we do have. By period 11, the simulation spent around $2800 to get $1100 in royalties, which is not far off from the actual figure of $3000 that was spent to get around $1000 of royalties. The simulation is optimistic, but this is no surprise as our RPC curve is usually more generous than the actual RPC. The output implies that even with consistent advertising, royalty payments for a single book will stagnate around the 12-period mark. All the while, revenue continues to drop every period as we continue investing in advertisement. So far, not so good.

But advertisements do not just generate income, they also generate a follower base. Not every consumer will become a follower, but those that do are much more likely to buy subsequent books in the series. Factoring in the lifetime value of a customer, the road for financial stability becomes clearer. Investing in and creating a *Follower Base* for the first book in a series will eventually lead to more sales later in the series. The more books there are in a series, the greater the lifetime value of each customer.

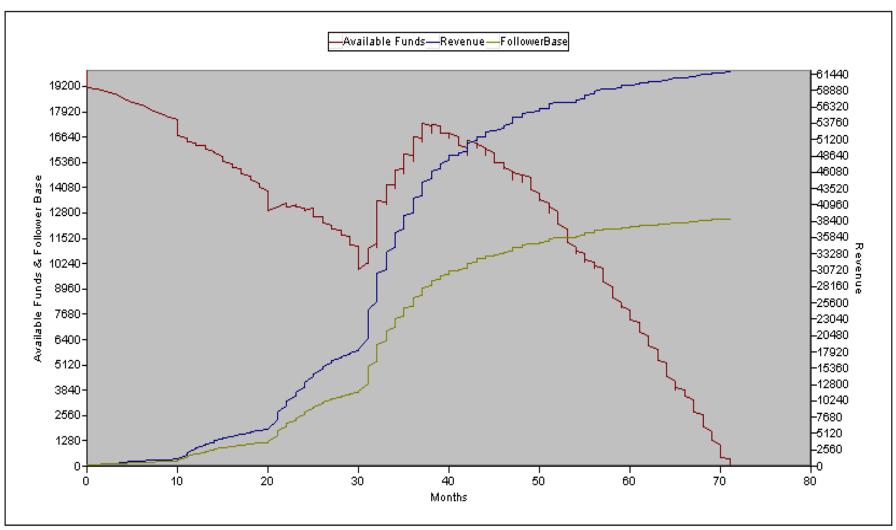
I took a simplified approach to working pre-existing followers into book sequels. The number of followers gained is proportional to the revenue from the same period. Every time a book is published, I increased the click potential by the number of current followers. This does not change anything for the first book, but every subsequent book starts off with more click potential. Like in the previous analysis, I also decided to reset the time counter for each book when a new book in the series is published. Intuitively, this is to mirror associated products (books, movies, etc.) getting attention again when a sequel is released (movie marathons, collector's bundles, etc.). Finally, according to this article that the company's founder shared with me, most authors can expect around 60% of a book's readers to read its sequel (read through rate). This makes marketing a series of books more efficient than if they were all standalone. For this simulation, I used a 60% read through rate for all books. Assuming there are four books already published, selling one book also means selling 60% ($.6^1$) of a second, 36% ($.6^2$) of a third, and 21.6% ($.6^3$) of a fourth.

Rerunning the model after these adjustments gives us the following new output. This time I tracked *Available Funds* and *Follower Base* on the left axis due to a similar scale between the two, and *Revenue* on the right. As before, I simulated until either 100 periods passed, or until *Available Funds* ran out. I also operated under the assumption that there were four books to this series, with each subsequent book taking 18 periods to write.



The key takeaway from this output is that once a customer base has been successfully established, a book series will become profitable. We see this inflection point around period 54. Which is coincides with the release of the fourth and last book in the simulated series. We then see profits for around ten periods, before the monthly decrease stagnates royalties again and drives *Available Funds* down to zero. This happens because we have no more books to release at this point.

It is interesting to note that this model implies that the timing between book releases is non-trivial. Shorter interarrival times between books lead to quicker profitability, but this does not always translate to more revenue and series health. It may be tempting to shorten that 54 month period of losses, but the benefit may not always outweigh the cost.

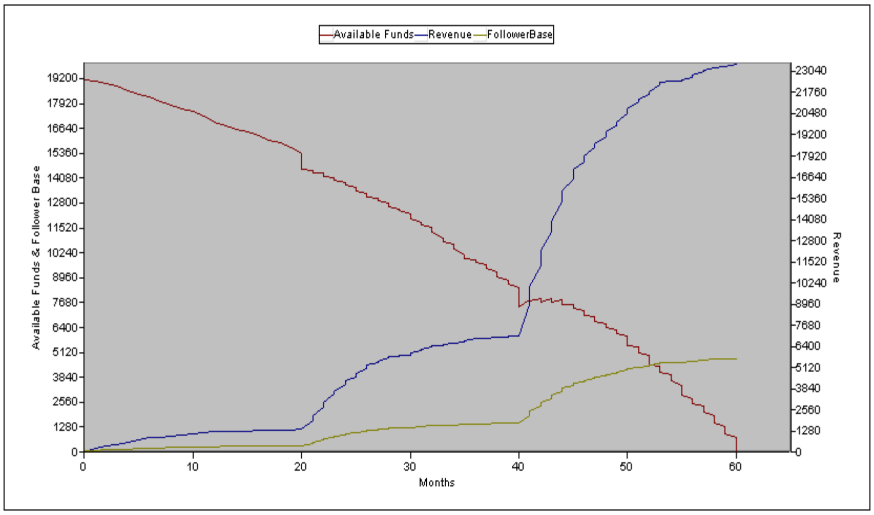


For example, keeping everything else *ceteris paribus* and changing only the time between publications, if we decreased the waiting period of 18 to a waiting period of 10, we instead get this output.

With this configuration, we see profits earlier around period 30 compared to period 54. However, we also run out of *Available Funds* faster (70 vs. 88).

Additionally, both *Follower Base* and *Total Revenue* are lower, even when compared to period 70 of the previous model (12.5k and $62k vs. 13k and $65k).

On the other hand, waiting too long also has negative effects to business viability. Increasing the waiting period to 20 (only 2 more months than 18) gives us this output.



This configuration does not ever become profitable and terminates at period 60. Moreover, the values at termination are nowhere near the previous two models'.

There seems to be a sweet spot in releasing books. Releasing them too close to each other cuts the attention lifespan of each individual book, and not all royalty sales are realized. This may be akin to market cannibalization. Releasing the books too far from each other runs the risk of falling out of relevancy and exhausting funds.

The length of a series has less impact on the overall behavior of the model than its publication rate. In general, the more books you plan to publish, the longer the total lifespan of the series. Exhausting funds prematurely overrides the behavior of the model regardless of how many books were planned.

The last thing of note is that the time to pay off the $2100 owed to the ghostwriter takes upwards of 54 months, even in the best model. This is not a realistic turnaround rate if we do not want angry ghostwriters complaining about not getting compensated. The only way to ameliorate this number in this current model, is by drastically increasing both publication rate and series length. This however sparks questions of feasibility and quality, and with no prior information on the sales behavior of quick turnaround books, it is unlikely that the current model would be accurate in dealing with this corner case. Therefore, a strategy of mass publication might not even successfully decrease author compensation turnaround time to acceptable levels.

Conclusions: After making and running this simulation, it makes sense why many business endeavors take time before becoming profitable. Profit often only comes with mass and scale, scale only comes with sufficient product exposure, and exposure only comes with time and investment. Anecdotally, this makes sense. I cannot remember getting into any book series in its infancy. When I started reading Harry Potter, the third book had already been released. I have also had similar experiences with Artemis Fowl and even the more recent Hunger Games. Another interesting result was that there was a sweet spot when it came to releasing books at realistic quantities. It surprised me that I had unwittingly simulated the effects of market cannibalization. One next step would be to do more research in the interaction between follower base and clicks. I believe that my current simulated interaction is still too naïve. Once sales data from the baseline book's sequel comes back, I can use this new data to retune the model.