An Empirical Perspective on Startup Valuations

Sebastian Quintero∗

Radicle Working Paper

Last Revised: January 30, 2019

Abstract

In this paper, we explore post-money valuations by venture capital stage classifications. We find that valuations can be described as being log-normally distributed, allowing us to think of them in terms of mean, median, and standard deviation, as well as build classical statistical models. We present an OLS log-log regression model as a tool to help practitioners approximate an undisclosed post-money valuation with considerable ease. Our results are statistically significant at the $p < 0.001$ level for all terms with an adjusted $R^2$ of 89 percent. We also describe the results of an equivalent Bayesian model to better understand the uncertainty of our OLS estimates.

1 Introduction

It’s often difficult to comprehend the significance of numbers thrown around in the startup economy. If a company raises a $550 million Series F round at a valuation of $4 billion [3]—how big is that really? How does that compare to other Series F rounds? Is that round approximately average when compared to historical financing events, or is it an anomaly?

At Radicle, a disruption research company, we use modern data science methods to better understand the entrepreneurial ecosystem. In our quest to remove opacity from the startup economy, we conducted an empirical study to better understand the underlying distribution of post-money valuations. While it’s popularly accepted that seed rounds tend to be at valuations somewhere in the $2m to the $10m range [18], there isn’t much data to back this up, nor is it clear what valuations really look like at subsequent financing stages. Looking back at historical events, however, we can see some anecdotally interesting similarities.

Google and Facebook, before they were household names, each raised Series A rounds with valuations of $98m and $100m, respectively. More recently, Instacart, the grocery delivery company, and Medium, the social publishing network, raised Series B rounds with valuations of $400m and $457m. Instagram wasn’t too dissimilar at that stage, with a Series B valuation of $500m before its acquisition by Facebook in 2012. Moving one step further, Square (NYSE: SQ), Shopify (NYSE: SHOP), and Wish, the e-commerce company that is mounting a challenge against Amazon, all raised Series C rounds with valuations of exactly $1 billion. Casper, the privately held direct-to-consumer startup disrupting the mattress industry, raised a similar Series C with a post-money valuation of $920m. Admittedly, these are probably only systematic similarities in hindsight because human minds are wired to see patterns even when there aren’t any, but that still makes us wonder—is there some underlying trend? Our research suggests that there is, but why is this important?

∗Head of Data Science, Inference and Machine Learning Group at Radicle Labs, Inc. Email: sebastian@rad.report.
I would like to thank Yuqing Luo, Nicholas Thorne, and Stuart Willson for thoughtful comments and suggestions. All errors are my own.
We think entrepreneurs, venture capitalists, and professionals working in corporate innovation or M&A would benefit greatly from having an empirical view of startup valuations. New company financings are announced on a daily cadence, and having more data-driven publicly available research helps everyone that engages with startups make better decisions. That said, this research is solely for informational purposes and our online tool is not a replacement for the intrinsic, from the ground up, valuation methods and tools already established by the venture capital community. Instead, we think of this body of research as complementary—removing information asymmetries and enabling more constructive conversations for decision-making around valuations.

This paper proceeds as follows. In the next section, we provide summary statistics as well as some thoughts on Crunchbase as a venture capital research database. We then outline our use of kernel density estimation to draw the probability density function for venture capital financing stages on a logarithmic scale. In Section 3 we present frequentist and Bayesian models that estimate the post-money valuation of a financing round by its stage classification and the amount of capital raised. We conclude with some comments on our use of this research in Startup Anomaly Detection\textsuperscript{TM}, our state of the art algorithm that estimates the probability that a startup will exit via an initial public offering or acquisition.

## 2 Making Sense of Post-Money Valuations

We obtained data for this analysis from Crunchbase, a venture capital database that aggregates funding events and associated meta-data about the entrepreneurial ecosystem. Our sample consists of 8,812 financing events since the year 2010 with publicly disclosed valuations and associated venture stage classifications. Table I below provides summary statistics\footnote{The sample size for the median amount of capital raised at each stage is much higher ($N=90k$) because round sizes are more frequently disclosed and publicly available.}.

<table>
<thead>
<tr>
<th>Stage</th>
<th>Mean</th>
<th>Median</th>
<th>SD</th>
<th>N</th>
<th>Median Capital Raised $N=91k$</th>
<th>Median Ownership Acquired $N=7,872$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angel</td>
<td>$4.7</td>
<td>$2.2</td>
<td>$19.9</td>
<td>1,267</td>
<td>$319k</td>
<td>10 percent</td>
</tr>
<tr>
<td>Pre-Seed</td>
<td>$3.4</td>
<td>$1.9</td>
<td>$9.8</td>
<td>162</td>
<td>$100k</td>
<td>8 percent</td>
</tr>
<tr>
<td>Seed</td>
<td>$4.6</td>
<td>$2.2</td>
<td>$12</td>
<td>5,297</td>
<td>$325k</td>
<td>10 percent</td>
</tr>
<tr>
<td>Series A</td>
<td>$280</td>
<td>$16</td>
<td>$1,872</td>
<td>757</td>
<td>$5m</td>
<td>20 percent</td>
</tr>
<tr>
<td>Series B</td>
<td>$850</td>
<td>$130</td>
<td>$3,451</td>
<td>409</td>
<td>$12m</td>
<td>16 percent</td>
</tr>
<tr>
<td>Series C</td>
<td>$1,048</td>
<td>$500</td>
<td>$2,504</td>
<td>322</td>
<td>$20m</td>
<td>13 percent</td>
</tr>
<tr>
<td>Series D</td>
<td>$1,432</td>
<td>$1,000</td>
<td>$2,103</td>
<td>266</td>
<td>$27m</td>
<td>12 percent</td>
</tr>
<tr>
<td>Series E</td>
<td>$2,881</td>
<td>$1,100</td>
<td>$6,864</td>
<td>175</td>
<td>$34m</td>
<td>10 percent</td>
</tr>
<tr>
<td>Series F</td>
<td>$3,975</td>
<td>$1,650</td>
<td>$7,676</td>
<td>89</td>
<td>$41m</td>
<td>9 percent</td>
</tr>
<tr>
<td>Series G</td>
<td>$5,167</td>
<td>$2,400</td>
<td>$9,834</td>
<td>43</td>
<td>$54m</td>
<td>8 percent</td>
</tr>
<tr>
<td>Series H</td>
<td>$8,656</td>
<td>$7,651</td>
<td>$10,620</td>
<td>20</td>
<td>$110m</td>
<td>6 percent</td>
</tr>
<tr>
<td>Series I</td>
<td>$9,789</td>
<td>$9,000</td>
<td>$9,910</td>
<td>5</td>
<td>$207m</td>
<td>5 percent</td>
</tr>
</tbody>
</table>

**Table I**

**Post-Money Valuation Summary Statistics (Millions, USD)**

*Data Source: Crunchbase*
The results of this paper depend on the suitability of Crunchbase as a research database. We don't expect our data samples to be complete because venture capital data never is—the majority of valuations are not disclosed—but can we reasonably believe that our samples are representative? Though the jury is still out by some researchers [10], we believe Crunchbase provides a sufficient view of the startup economy because we’ve found it to be on par with long-standing research databases after the year 2010, and often better the closer we move towards the present\(^2\). To inspect the quality of the data, however, we calculated the percentage ownership acquired for each observation by dividing the amount of capital raised by the post-money valuation [19]. The median percentage ownership acquired starts between 8 and 10 percent for angel and pre/seed rounds, peaks at 20 percent for Series A, and then declines progressively to a low of 5 percent at the Series I stage as valuations skyrocket. It’s commonly believed that venture capitalists insist on obtaining 20 percent ownership for Series A rounds [24], which suggests that our early stage samples are representative. The decreasing ownership rate reflects the intuition that investors buy increasingly smaller stakes in startups as they invest increasingly larger levels of capital, but at such high valuations that those smaller stakes would yield very handsome returns should the company continue to grow and have a successful liquidity event. Previous research on startup failure between financing stages further suggests that late-stage companies are exceptionally rare, so the samples are representative of what are in reality very small populations [16]. That said, it’s reasonable to believe that valuation samples from any venture capital database would tend to be left-skewed because founders and investors are only reasonably incentivized to disclose the details of a round if they are deemed worthy of a press release.

To better understand the nature of post-money valuations, we assessed their distributional properties using kernel density estimation (KDE), a non-parametric approach commonly used to approximate the probability density function (PDF) of a continuous random variable [8]. Put simply, KDE draws the distribution for a variable of interest by analyzing the frequency of events much like a histogram does. Non-parametric is just a fancy way of saying that the method does not make any assumption about the data being normally distributed, which makes it perfect for exercises where we want to draw a probability distribution but have no prior knowledge about what it actually looks like. Figures I and II show the probability density functions for venture capital stages on a logarithmic scale, with vertical lines indicating the median for each class.

Why on a logarithmic scale? Post-money valuations are power-law distributed\(^3\), as most things are in the venture capital domain [5], which means that the majority of valuations are at low values but there’s a long tail of rare but exceptionally high valuation events. However, post-money valuations can also be described as being log-normally distributed\(^4\), which means the natural logarithm of valuations is normally distributed. Series A, B, and C valuations may be argued as being bimodal log-normal distributions, and seed valuations may be approaching multimodality (more on that later), but technical fuss aside, this detail is important because log-normal distributions are easy for us to understand using the common language of mean, median, and standard deviation—even if we have to exponentiate the terms to put them in dollar signs. More importantly, this allows us to consider classical statistical methods that assume normality.

Founders that seek venture capital to get their company off the ground usually start by raising an angel or a seed round. An angel round consists of capital raised from their friends, family members, or wealthy individuals, while seed rounds are usually a startup’s first round of capital from institutional investors [18]. The median valuation for both angel and seed is $2.2m USD, while the

---

\(^2\)We purchase Crunchbase data and do not receive a rate discount for saying this.

\(^3\)See Table I.

\(^4\)KDE is generally sensitive to the bandwidth selection hyper-parameter, but we saw no real difference between Scott’s and Silverman’s ”Rule of Thumb” bandwidth options and used Scott’s Factor for the final estimates [20].
median valuation for pre-seed is $1.9m USD. While we anticipated some overlap between angel, pre-
seed and seed valuations, we were surprised to find that the distributions for these three classes of 
rounds almost completely overlap. This implies that these early stage classifications are remarkably 
similar in reality. That said, we think it’s possible that the angel sample is biased towards the larger 
events that get reported, so we remain slightly skeptical of the overlap. And as mentioned earlier, 
the distribution of seed stage valuations appears to be approaching multimodality, meaning it 
has multiple modes. This may be due to the changing definition of a seed round and the recent 
institutionalization of pre-seed rounds, which are equal to or less than $1m in total capital raised and 
have only recently started being classified as ‘Pre-seed” in Crunchbase. There’s also a clear mode 
in the seed valuation distribution around $7m USD, which overlaps with the Series A distribution, 
suggesting, as others recently have, that some subset of seed rounds are being pushed further out 
and resemble what Series A rounds were 10 years ago [1].

Approximately 21 percent of seed stage companies move on to raise a Series A [16] about 18 
months after raising their seed—with approximately 50 percent of those Series A companies moving 
on to a Series B a further 18-21 months out [17]. In that time the median valuation jumps to $16m 
at the Series A and leaps to $130m at the Series B stage. Valuations climb further to a median 
of $500m at Series C. In general, we think it’s interesting to see the binomial nature as well as 
the extent of overlap between the Series A, B, and C valuation distributions. It’s possible that the 
overlap stems from changes in investor behavior, with the general size and valuation at each stage 
continuously redefined. Just like some proportion of seed rounds today are what Series A rounds 
were 10 years ago, the data suggests, for instance, that some proportion of Series B rounds today 
are what Series C rounds used to be. This was further corroborated when we segmented the data
by decades going back to the year 2000 and compared the resulting distributions. We would note, however, that the changes are very gradual, and not as sensational as is often reported [12].

The median valuation for startups reaches $1b between the Series D and E stages, and $1.65 billion at Series F. This answers our original question, putting Peloton’s $4 billion-dollar appraisal at approximately the 81st percentile of valuations at the Series F stage, far above the median, and indeed above the median $2.4b valuation for Series G companies. From there we see a considerable leap to the median Series H and I valuations of $7.7b and $9b, respectively. The Series I distribution has a noticeably lower peak in density and higher variance due to a smaller sample size. We know companies rarely make it that far, so that’s expected. Lyft and SpaceX, at valuations of $15b and $27b, respectively, are recent examples of companies that have made to the Series I stage\(^5\).

We classified each stage into higher level classes using the distributions above, as one of Early (Angel, Seed), Growth (Series A, B, C), Late (Series D, E, F, G), or Private IPO (Series H, I). With these aggregate classifications, we further investigated how valuations have faired across time and found that the medians (and means) have been more or less stable on a logarithmic scale\(^6\). What has changed, since 2013, is the appearance of the ”Private IPO” [11, 13]. These rounds, described above with companies such as SpaceX, Lyft, and others such as Palantir Technologies, are occurring later and at higher valuations than have previously existed. These late-stage private rounds are at such high valuations that future IPOs, if they occur, may be down rounds [22].

\(^5\) In December 2018 SpaceX raised a Series J round, which is a classification not analyzed in this paper.

\(^6\) See Figure III.
3 Approximating an Undisclosed Valuation

Since the valuation time series are stationary when logged, and come from more or less well-behaved log-normal distributions, we designed a model to predict a round’s post-money valuation by its stage classification and the amount of capital raised. Why might this be useful? Well, the relationship between capital raised and post-money valuation is true by definition, so we’re not interested in establishing a statistical relationship in the classical sense. A startup’s post-money valuation is equal to an intrinsic pre-money valuation calculated by investors at the time of investment plus the amount of capital raised [19, 21]. However, pre-money valuations are often not disclosed, so a statistical model for estimating an undisclosed valuation would be helpful when the size of a round is available and its stage is either disclosed as well or easily inferred. We formulated an ordinary least squares log-log regression model after considering that we did not have enough stage classifications and complete observations at each stage for multilevel modeling and that it would be desirable to build a model that could be easily understood and utilized by founders, investors, executives, and analysts. Formally, our model is of the form:

$$\log(y_i) = \beta_0 + \sum_{k=1}^{11} \beta_k \log(c_i) \cdot r_{ik} + \epsilon_i$$

where $y_i$ is the output post-money valuation, $c_i$ is the amount of capital raised, $r_i$ is a binary term that indicates the financing stage, and $\epsilon_i$ is the error term. $\log(c_i) \cdot r_i$ is therefore an interaction term that specifies the amount of capital raised at a specific stage. The model we present does not
include stage main effects because the model remains the same, whether they’re left in or pulled out, while the coefficients become reparameterizations of the original estimates [23]. In other words, boolean stage main effects adjust the constant and coefficients while maintaining equivalent summed values—increasing the mental gymnastics required for interpretation without adding any statistical power to the regression. Capital main effects are not included because domain knowledge and the distributions above suggest that financing events are always indicative of a company’s stage, so the effect is not fixed, and therefore including capital by itself results in a misspecified model alongside interaction terms. As is standard practice, we used heteroscedasticity robust standard errors to estimate the $\beta$ coefficients, and residual analysis via a fitted values versus residuals plot confirms that the model validates the general assumptions of ordinary least squares regression. There is no multicollinearity between the variables, and a Q-Q plot further confirmed that the data is log-normally distributed. The results are statistically significant at the $p < 0.001$ level for all terms with an adjusted $R^2$ of 89 percent and an F-Statistic of 5,900 ($p < 0.001$). Table II outlines the results. Monetary values in the model are specified in millions, USD.

### Table II
Log-Log Regression Results

<table>
<thead>
<tr>
<th>Covariate</th>
<th>$\beta$ Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>1.7910***</td>
<td>(0.016)</td>
</tr>
<tr>
<td>Angel</td>
<td>0.6115***</td>
<td>(0.020)</td>
</tr>
<tr>
<td>Pre-Seed</td>
<td>0.6609***</td>
<td>(0.034)</td>
</tr>
<tr>
<td>Seed</td>
<td>0.6445***</td>
<td>(0.011)</td>
</tr>
<tr>
<td>Series A</td>
<td>0.9650***</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Series B</td>
<td>1.0177***</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Series C</td>
<td>1.0536***</td>
<td>(0.010)</td>
</tr>
<tr>
<td>Series D</td>
<td>1.0707***</td>
<td>(0.008)</td>
</tr>
<tr>
<td>Series E</td>
<td>1.1066***</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Series F</td>
<td>1.1209***</td>
<td>(0.017)</td>
</tr>
<tr>
<td>Series G</td>
<td>1.1049***</td>
<td>(0.023)</td>
</tr>
<tr>
<td>Series H</td>
<td>1.1569***</td>
<td>(0.041)</td>
</tr>
<tr>
<td>Series I</td>
<td>1.2371***</td>
<td>(0.106)</td>
</tr>
<tr>
<td>N</td>
<td>7,862</td>
<td></td>
</tr>
<tr>
<td>Adj-R$^2$</td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>F-Statistic</td>
<td>5,900</td>
<td></td>
</tr>
<tr>
<td>Prob (F-statistic)</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

*Notes: Heteroscedasticity robust standard errors in parentheses.*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

The model can be interpreted by solving for $y$ and differentiating with respect to $x$ to get the marginal effect. Therefore, we can think of percentage increases in $x$ as leading to some percentage increase in $y$. At the seed stage, for example, for a 10 percent increase in money raised a company can expect a 6.4 percent increase in their post-money valuation, ceteris paribus. That premium increases as a company continues through the venture capital funnel, peaking at the Series I stage with a 12.4 percent increase in valuation per 10 percent increase in capital raised.

7 Of course, whether or not a stage classification is agreed upon by investors and founders and specified on the term sheet is another matter.
In practice, an analyst could approximate an unknown post-money valuation by specifying the amount of capital raised at the appropriate stage in the model, exponentiating the constant and the $\beta_k$ term, and multiplying the values, such that:

$$e^{\beta_0} \cdot c_i^{\beta_k} \approx y_i$$

$$e^{1.791} \cdot 2^{0.6445} \approx 9.4$$

$$e^{1.791} \cdot 35^{1.0177} \approx 223.5$$

$$e^{1.791} \cdot 200^{1.0707} \approx 1744$$

Using the first equation and the values in Table II, the estimated undisclosed post-money valuation of a startup after a $2m seed round is approximately $9.4m USD—for a $35m Series B, it’s $224m—and for a $200m Series D, it’s $1.7b. Subtracting the amount of capital raised from the estimated post-money valuation would yield an estimated pre-money valuation.

Can it really be that simple? Well, that depends entirely on your use case. If you want to approximate a valuation and don’t have the tools to do so, and can’t get on the phone with the founders of the company, then the calculations above should be good enough for that purpose. If instead you’re interested in purchasing a company, this is a good starting point for discussions, but you probably want to use other valuation methods, too. As mentioned earlier, this research is not meant to supplant existing valuation methodologies established by the venture capital community.

Figure IV. Scatter plot and line of best fit for individual regressions.

As far as estimation errors, you can infer from the scatter plot above that, for the predictions at the early stages, you can expect valuations to be off by a few million dollars—for growth-stage companies, a few hundred million—and in the late and private IPO stages, being off by a few
billion would be reasonable. Of course, the accuracy of any prediction depends on the reliability of the estimated means, i.e., the credible intervals of the posterior distributions under a Bayesian probabilistic framework, as well as the size of the error term from omitted variable bias—which is not insignificant.

We can reformulate our model in a directly comparable probabilistic Bayesian framework, in vector notation, as:

\[ \log(y) | \beta, \sigma, X \sim \mathcal{N}(X\beta, \sigma^2 I) \]

where the distribution of \( \log(y) \) given \( X \), an \( n \times k \) matrix of interaction terms, is normal with a mean that is a linear function of \( X \), observation errors are independent and of equal variance, and \( I \) represents an \( n \times n \) identity matrix [6]. We fit the model with a noninformative flat prior using the No-U-Turn Sampler (NUTS), an extension of the Hamiltonian Monte Carlo MCMC algorithm [9], for which our model converges appropriately and has the desirable hairy caterpillar sampling properties.

![Figure V. Bayesian 95 percent credible intervals of the posterior distributions.](source)

The 95 percent credible intervals in Figure V suggest that posterior distributions from angel to series E, excluding pre-seed, have stable ranges of highly probable values around our original OLS coefficients. However, the distributions become more uncertain at the later stages, particularly for series F, G, H, and I. This should be obvious, considering our original sample sizes for the pre-seed class and for the later stages. Since the data needs to be transformed back to its original scale for appropriate estimation, and the fact that the magnitudes of late stage rounds tend to be very high, such changes in the exponential will lead to dramatically different prediction results. As with any simple tool then, your mileage may vary. For more accurate and precise estimates,
we’d suggest hiring a data scientist to build a more sophisticated machine learning algorithm or Bayesian model to account for more features and hierarchy. If your budget doesn’t allow for it, the simple calculation using the estimates in Table II will get you in the ballpark.

4 Concluding Remarks

This paper provides an empirical foundation for how to think about startup valuations and introduces a statistical model as a simple tool to help practitioners working in venture capital approximate an undisclosed post-money valuation. That said, the information in this paper is not investment advice, and is provided solely for educational purposes from sources believed to be reliable. Historical data is a great indicator but never a guarantee of the future, and statistical models are never correct—only useful [2]. This paper also makes no comment on whether current valuation practices result in accurate representations of a startup’s fair market value, as that is an entirely separate discussion [7].

This research may also serve as a starting point for others to pursue their own applied machine learning research. We translated the model presented in this article into a more powerful learning algorithm [8] with more features that helps us fill-in the missing post-money valuations in our own database. These estimates are then passed to Startup Anomaly Detection™, an algorithm we’ve developed to estimate the plausibility that a venture backed startup will have a liquidity event given the current state of knowledge about them. Our machine learning system appears to have some similarities with others recently disclosed by GV [15], Google’s venture capital arm, and Social Capital [14], with the exception that our probability estimates are available as part of Radicle’s research products.

Companies will likely continue raising even later and larger rounds in the coming years, and valuations at each stage may continue being redefined, but now we have a statistical perspective on valuations as well as greater insight into their distributional properties, which gives us a foundation for understanding disruption as we look forward.
References


[22] Tunguz, Thomasz. "The Runaway Train of Late Stage Fundraising", 1 April, 2015. Article
