The new risk and return of venture capital

François Burguet[1], Loïc Maréchal[1,2], and Alain Mermoud[1]

[1]Cyber-Defence Campus, armasuisse Science and Technology
[2]Department of Information Systems, HEC Lausanne, University of Lausanne

November 7, 2022

Abstract

This paper revisits the study of Cochrane (2005), to estimate the risk and returns of venture capital investments while correcting for the selection bias. We use an up-to-date dataset and enhance it to account for missing firm valuations using machine learning. The model is able to infer, with a median error of less than 4%, the true log value of the firm, for a total of nearly 120,000 observations, or six times more than the original paper, from 2010 to 2022. We find an annualized expected return of around 38%, an annualized alpha of 32.14%, a beta of 1.37, and a 40% idiosyncratic risk. Our results are robust to the choice of the benchmark index. Depending on the sector, we find a beta lower than 1 for the health industry and of up to 1.86 for the tech sector. The health industry exhibits the lowest alpha (24%) and the tech the highest (36%). We use the cyber-security sector as a case study and find an alpha of 36%, on par with the tech sector, but with a lower beta of 1.56.

JEL classification: C01, C14, C51, G12, G24

Keywords: private equity, venture capital, machine learning

1. Introduction

The main focus of this work is to assess the risk and return, modeled by the alpha, beta, and sigma, of venture capital investments. Theses parameters are fundamental variables in market analysis, as they allow to assess the behavior of an investment in simple terms and to compare two investments. The alpha (α) is the excess return on an investment after adjusting for market risk. Beta (β) is a measure of volatility relative to a benchmark, such as the S&P 500 index. It is a measure of the systematic risk of the investment, *i.e.* risk that is not specific to a particular project (such as operational risk). Sigma (σ) is the volatility of the returns of the investment over a certain period of time. These variables are widely used in public equity market analyses, but are much harder to estimate when dealing with private equity markets for multiple reasons. First, the price process is not continuous, and neither is the return process. We observe very little valuations, as they occur only after a financing round, and are rarely publicly disclosed. Second, observations are asynchronous. As stated before, we observe prices at random intervals, making it impossible to compute cross-correlations. Finally, the private equity market is opaque by nature. Compared to public markets, very little data is available and thus is a hurdle for any systematic research.

In this paper we adapt and apply to a new dataset a maximum likelihood estimation (MLE) method to estimate a set of parameters using a log-normal return model. Our research is primarily inspired by previous work from Cochrane (2005). The model assumptions are backed by the data, especially the probability distribution of exit and the distribution of log-returns, which we find to follow a logistic and log-normal distribution respectively. Since we only observe valuations when a firm gets a new financing round, is acquired, or exits, we observe valuations of successful companies more often. This bias should not be underestimated as it could significantly bias results upwards. To overcome this issue, we add a selection bias and error measurement protections to the model. Returns on private equity (PE) investments are calculated from financing rounds to exit, not from round to round, as there is no concrete way to access liquidity between rounds. In order to alleviate the data sparsity issue, we come up with a machine-learning approach to estimate missing post-money valuations (PMV) after a financing round. We multiply by a factor of five the number of usable observations for a total of roughly 120,000 data points. The model was able to accurately predict log-PMVs used for the log-returns in the MLE procedure with a median percentage error of about 3%.

For the global PE market, we find an annualized arithmetic α of 32.14% for the S&P500 (total returns), 32.65% for the NASDAQ and 33.21% for the RUSSELL 2000. All benchmarks yield very large positive alphas, in line with previous findings. We also find β between 1.11

(RUSSELL 2000) and 1.37 (S&P500 TR), which indicate that PE investments are riskier than those made in public equity markets, even when compared to small stocks indices. These results are once again consistent, as the RUSSELL 2000 yields the lowest β. Expected arithmetic returns are surprisingly high, at around 38% annualized. These high expected returns are primarily due to the very high estimated annualized total risk σ at almost 46% for all benchmarks. "Venture capital investments are like options; they have a small chance of a huge payoff".[1]

We analyze specific sectors of the industry. Our findings are that the tech and retail industries are the most risky sectors of the market with β up to 1.86, but also the more valuable with the highest annualized α at 36%. On the contrary, the health industry and the rest of the market (neither of the three previous sectors) are by far less risky, with β around 0.80 down to 0.64. The health industry exhibits the "lowest" α between 24% and 26%, which is still much higher than for other asset classes. Finally, for the cyber-security industry, we find a more nuanced total volatility σ between 41% and 49%, β accordingly lower between 1.16 and 1.56 and high values for α, between 32.48 up to 36% on the S&P500 index. The cyber-security industry seems to outperform other sectors in the PE market, while being less risky. The remainder of this thesis is organized as follows. Section 2 reviews the literature and develop related hypothesis. Section 3 details the data and the methodology. Section 4 presents the results and Section 6 concludes.

2. Literature review and hypothesis development

2.1. Private equity market

Studies in private equity markets, face two problems compared to those in public equity market. First, as opposed to public companies, private firms are not required by law to disclose to the public any of their financial statements. In the U.S. public equity this is done with the Securities and Exchange Commission (SEC) forms 10-Q (quarterly) or 10-K (annually). In addition to these forms, public firms must send an annual report to their shareholders. None of this is mandatory for private firms. This makes it a lot more difficult to estimate their past performance and future prospects, without insider information. Second, they have no public shares to trade, thus we do not observe any (quasi-)continuous price process or market capitalization. Almost none of the estimations employed in research on public equity are usable in this context. Nonetheless, private firms still issue shares to investors in various forms, with different optionality clauses.

[1]The risk and return of Venture Capital, John H. Cochrane (2005)

Private equity analysts rely on insider or private information to evaluate projects. The common practice is to value a company after (and/or before) a financing event. A financing event is any type of event during which the firm receives equity. The most common is the financing round to collect cash. Other types of financing events include emission of notes (loans), debt financing, non-equity assistance (furniture or real estate), grants and more. This valuation is called pre- (post-)money valuation when it is done before (after) the financing event. Again, different metrics are used in the industry but the most common one consists in multiplying the per-share price of the most recent event by the fully diluted number of common shares. This simplified calculation does not account for the optionality of the investment contract and assumes that all shares have the same value, regardless of their type (common or preferred shares or convertibles notes). This matter is extensively investigated by Gornall and Strebulaev (2020).

2.2. Venture capital risk and return

The scope of this work is to update and extend Cochrane (2005). A number of studies have tried to overcome the challenges of evaluating risk and returns in venture capital. In his seminal study, Cochrane (2005) uses a maximum likelihood estimation method to obtain, the values of α, β, and σ^2 for the private equity market. He simulates and analyzes the market as a whole but also particular sectors such as healthcare and biotech, tech companies, and retail services. He finds a mean arithmetic return of 59%, an alpha of 32%, a beta of 1.9 and a volatility of 86% (which corresponds to 4.7% daily volatility). Because of the returns distribution, which is heavily positively skewed, he fits a logarithmic model. The outstanding feature of this research is the selection bias correction. As most of PE data is private and hard to gather, successful firms are more often observed with good data than small and underperforming companies. He treats this bias using a simple probabilistic approach, trimming detected outliers from the simulation.

Ewens (2009) updates Cochrane's methodology, but focuses on round-to-round returns. He uses the logarithmic model, but splits it into a three-regime mixture model (failure, medium returns, and "home-runs"), and a separate holding period model, to simulate the time between two financing events. After correction for the selection bias, he obtains an alpha of 27% and a beta of 2.4, values that are lower and higher than Cochrane's, respectively. He finds that 60% of all venture capital investments have a negative mean log return and substantial idiosyncratic volatility. Although different from Cochrane's results, the underlying conclusions remain similar. Namely, private equity investments exhibits positive alpha, large beta, and a high volatility in arithmetic returns due to the idiosyncratic risk of

individual projects.

Korteweg and Nagel (2016) extend the popular public market equivalent (PME) method to assess private equity funds. They estimate monthly arithmetic alphas of 3.5%, in line with those of Cochrane (2005) and Korteweg and Sorensen (2010). They argue that their method delivers similar results in a "much simpler and more robust fashion that does not require specific distributional assumptions and rather cumbersome estimation of a selection model".

Moskowitz and Vissing-Jørgensen (2002) obtain more nuanced results and conclude that returns of private equity are not higher than those on public equity. Their estimates suggest that the index of private equity is likely as volatile as the public equity index and that aggregate private equity returns are highly correlated with those of public equity. By finding a high idiosyncratic risk of single private firms, they conclude that the aggregate return overestimates the average returns to investors.

In a second line of research, Axelson and Martinovic (2015) and Franzoni, Nowak, and Phalippou (2012) estimate abnormal returns and risk factor loadings with standard regression techniques by using either internal rates of return (IRRs) or modified internal rates of return (MIRRs). Driessen, Lin, and Phalippou (2012) and Ang, Chen, Goetzmann, and Phalippou (2018) present an approach that extends the standard internal rate of return (IRR) calculation to a dynamic setting in which they solve for the abnormal returns and risk exposures using the Generalized Method of Moments (GMM). This approach requires only a cross-section of observable investment cash flows. They both identify parameters by using a net present value (NPV) framework. Beta coefficients reported by Korteweg and Sorensen (2010) and Driessen et al. (2012) both average to 2.8.

The literature is very sparse on the topic, and there is no consensus on methodology and results. Nonetheless, many of the previous work tend to estimate similar values (see Cochrane, 2005; Ewens, 2009; Korteweg and Nagel, 2016. Cochrane's methodology stands out by its simplicity and has been successfully used in later research (see, *e.g.*, Gornall and Strebulaev, 2020). It also focuses on returns from rounds to IPO, which is not the case for Korteweg and Sorensen (2010). For these reasons, we choose to review and update his work with the most recent data available.

2.3. Indices and benchmarks

Directly related to private equity risk and return research, several works have been carried on indexing and benchmarking the private equity market. Peng (2001) builds a venture capital index spanning the 1987–1999 period that consists of almost 13,000 financing rounds

in more than 5,643 firms. He addresses the three common problems with private equity data (missing data, censored data, and sample selection) using a re-weighting procedure and method of moment regressions. He finds high and volatile returns to venture capital (geometric average return is 55.18% per year). His venture capital index has a much higher volatility than the S&P 500 and NASDAQ indices, but is substantially correlated to the latter. Its index betas with those of the S&P500 and NASDAQ are 2.4 and 4.7, respectively.

Hwang, Quigley, and Woodward (2005) build an index of value for venture capital using Sand Hill Econometrics data. They overcome the lack of pricing data using a repeat valuation model based on the same principle as Peng (2001). Their index has a beta of 0.03 with the S&P500, an alpha of 103%, and a beta of 0.4 and alpha of 59% for the NASDAQ.

Schmidt (2006) studies the benefits of including private equity assets in portfolios, using CEPRES' Private Equity Analyzer. By simulating investments in individual benchmark stocks with the same timing, he observes exact benchmark performances, measured by the internal rate of return (IRR). His optimal mixed-asset portfolio consist of 3% to 65% private equity assets, for an ideal portfolio size between 20 and 28 investments.

Cumming, Haß, and Schweizer (2013) show that none of the three typical indices (listed private equity, transaction-based private equity, or appraisal value-based private equity indices) is fully suitable for portfolio optimization. They introduce a new monthly benchmark index that achieves superior quantitative results. Namely, they achieve higher Sharpe ratios and lower risk in portfolios where the benchmark is used.

2.4. Venture capital fund performance

This work also fits into the broader literature on venture capital funds performance. Kwon, Lowry, and Yiming (2020) investigate the factors influencing the increasing trend of mutual fund investments in private firms. They conclude that mutual fund investments enable companies to stay private longer, contribute toward higher abnormal returns and are associated with higher IPO allocations when the firms go public.

Chernenko, Lerner, and Zeng (2021) investigates the impact of new investors, specifically mutual funds, on the governance of entrepreneurial firms. Mutual funds' liquidity concerns can create a wedge between their incentives and those of earlier-stage venture investors, which affect the contracts between entrepreneurs and investors. They conclude that mutual funds are more likely to invest in late rounds, hot sectors, and larger firms and that larger mutual funds and those with less volatile fund flows are more likely to invest in unicorns.

Phalippou (2009) study the overstated performance of private equity funds and conclude that a large part of performance is driven by inflated accounting valuation of ongoing invest-

ments. They find an average net-of-fees fund performance of 3% per year below that of the S&P 500 and underperformance drops to 6% when adjusting for risk.

Driessen et al. (2012) develop a new methodology to estimate abnormal performance and risk exposure of non-traded assets from cash flows. They find a high market (S&P500) beta (2.73), a low alpha (-1.09% monthly) and underperformance before and after fees for venture capital funds. When compared to Fama-French 3-factor benchmark, they find larger alpha (-0.74%) and conclude that VC returns resemble those of small growth stocks.

Harris, Jenkinson, and Kaplan (2016) use cash flow data derived from the holdings of close to 300 institutional investors and find that average buyout fund returns before 2006 have exceeded those from public markets; averaging about 3% to 4% annually and that post-2005 year returns have been roughly equal to those of public markets.

The sparse literature over venture capital risk and returns show room for great improvement. Most notable results are dated back to mid 2000s. Although there is no consensus on the results and methodology, several works converge towards similar estimates, such as Cochrane (2005), Ewens (2009) or Korteweg and Nagel (2016). Using Cochrane's most straightforward approach, we investigate how these results have evolved using up-to-date data on all industries and how they compare across industries.

3. Data and methodology

3.1. Data

3.1.1. Crunchbase

Crunchbase is a commercial database that provides access to financial and managerial data on private and public companies globally. It was created in 2007 by TechCrunch, an American online newspaper focusing on high tech and startup companies. Since 2015 it is maintained by Crunchbase Inc., a "Data as a Service" firm founded in 2015, located in San Francisco, California, and hiring 153 employees (as of 2019). Although recent, this database has been largely adopted by both academics (Besten *den*, 2021) and industry practitioners. It is also used by international organizations such as the OECD (see Dalle, Besten *den*, and Menon, 2017). Finally, no previous academic work has been carried out on the topic of this thesis to our knowledge.

Crunchbase collects data following a multifaceted approach, combining crowd-sourcing (either through venture programs or direct community contributions), machine learning (monitoring top news publications to capture every notable funding round, acquisition, and exit), in-house processing (to verify data integrity and accuracy), or relying on third-party

providers for additional metrics (such as company valuation, interest signals, or mobile apps analytics). Crunchbase updates and revises data on a daily basis, which is organized into several entities, the main ones being:

- Organizations: this set of entities contains administrative information on private and public companies, investment funds, or institutions. It includes information about businesses, contact details, description, social media links, last funding round, geographic location, or number of employees.
- People: this set of entities contains information about physical persons such as investors or CEOs. It includes age, CV, degrees, social media links, number of organizations founded, gender, job title, and rank (algorithmic rank assigned to the top 100,000 most active people).
- Events: this set contains information on the type of event (meetup, hackathon, conference, or festival), dates, number of participants, organizers information, sponsors, and location.
- Funding Rounds: this set includes all the funding rounds registered by Crunchbase. It contains over 200,000 entries. The fields include funded organization, investors, date, type of round (seed, series A, B, C, ..., or debt issuing), amount of money raised, number of investors, funding round rank (algorithmic rank assigned to the top 100,000 most active funding rounds), target money raised, and pre-/post-money valuations.
- Acquisitions: this set includes all the acquisitions registered by Crunchbase. Includes name of the acquiree, name of the acquirer, organization locations, last funding rounds, revenue range, amount of acquisition, acquisition type (acqui-hiring, acquisition, leverage buyout (LBO), management buyout, or merger), date of announcement, date of completion, acquisition terms (cash, stock, or both).
- Initial Public Offerings (IPOs): all the IPOs (listing or delisting) registered by Crunchbase. It includes the amount raised during IPO, date of the event, share price, number of shares outstanding, number of shares sold, stock exchange where the IPO took place, and stock symbol.

For most of the entities, a large number of observations are missing, regardless of the information confidentiality. In particular, post money valuations (PMVs) and IPO share prices, two essential variables for this study are often not available. PMV is an accounting estimates provided by a VC firm and hence, this restricted number of observations is consistent with this type of data source. On the other hand, the share price of an IPO is public information and provided by the regulation authority of each country (*e.g.* the SEC in the United States). However, Crunchbase does not consistently provide values for this field. We discuss that issue and the solutions below.

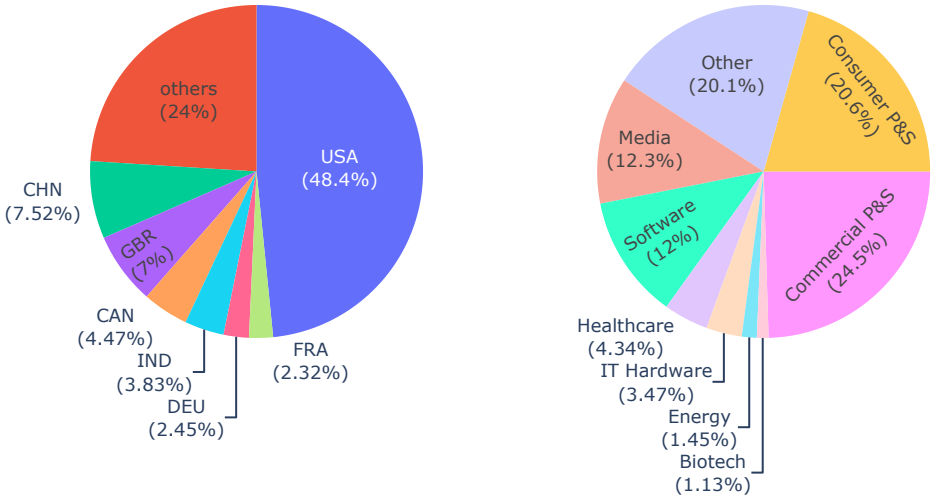

Fig. 1: Distribution of funding rounds across main countries and sectors

The data provided is an aggregate from many sources, without a clearly defined coverage. This may induce heterogeneity, and the quantity and quality of observations is likely to vary across country, industry, or period. For instance, U.S. companies are over-represented compared to other nations (China coming second), as shown in Figure 1. Moreover, prior to the 2000s, we observe very few data points, and the amount of data has been growing exponentially since then.

Even though the data set rapidly expands, covering close to all industrial sectors, the primary focus of Crunchbase is the technology industry. This is a critical aspect to consider when using Crunchbase as a primary source of data, since it it is not representative of the economy. Figure 1 shows the distribution of funding rounds data per industrial sector.

We download data from May 2022 and nearly 88% of the recorded funding rounds take place after 2010 (429,930 over 488,861 total rounds). Before 2013, the number of observations is of 21 per day. After 2013, this number jumps to 142 per day, and the trend is growing. Although the VC market considerably expanded since the 2008 financial crisis, Crunchbase arguably lacks observations before 2013, compared to today. This can introduce a large bias, as the post internet-bubble area has been a decisive period for numerous companies and VC funds. Nonetheless, this was a period of economical turmoil where a lot of companies either closed, or boomed. Thus, indicators for this period should also reflect this unique context, and it may not be adequate to include them in a more contemporary analysis. We further investigate this matter in Section 4

Next, we analyze the data distribution with respect to the type of funding, and the corresponding stage (early, mid, or late). Crunchbase uses several labels for each rounds and thus, we allocate them to one of the three stages. For early stage rounds, we include

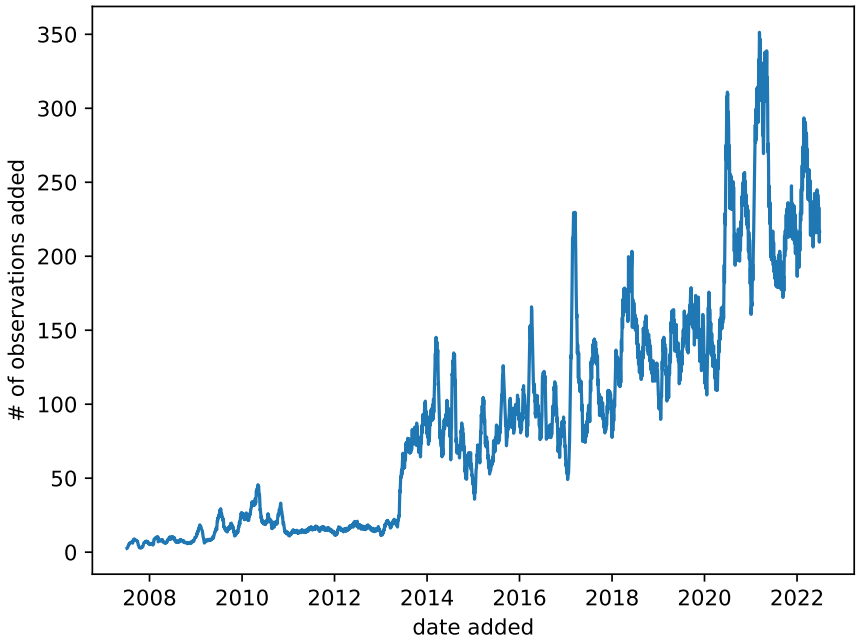

Fig. 2: Number of daily added observations (30 days moving average)

pre-seed, seed, angel rounds, products and equity crowdfunding, ICOs (initial coin offering), and non-equity assistance (like furniture, offices, or machinery). For mid-stage rounds we include series A and B, as well as corporate rounds (when a company invests in another company). We classify corporate rounds as mid-stage, because the median raised amount and median post-money valuations are closer to what we observe for Series B than for Series C. Lastly, for late rounds, we include all rounds starting from series C. All other types of rounds are discarded because there is no clear cut for their categorization. This includes post-IPO rounds, unknown or undisclosed series, debt financing, grants, secondary market sells, and convertible notes. Eventually, we use the complete dataset to estimate the model, and this preliminary analysis is only meant to give an overview of the dataset. In his study, Cochrane (2005) gathers data from VentureOne, which records a financing round only if a VC firm with more than $20 million asset under management is involved in the round. Conversely, Crunchbase records almost every deal that occurs in the VC market, using the above-mentioned techniques. In that regard, the Crunchbase dataset seems less impacted by the survivor bias.

67% of the Crunchbase observations are early stage rounds (210,962 points). The mid-stage rounds account for 27% of the sample (84,197 points) and late stage rounds only account for 7% of the total number of classified rounds (20,423 points). Although class

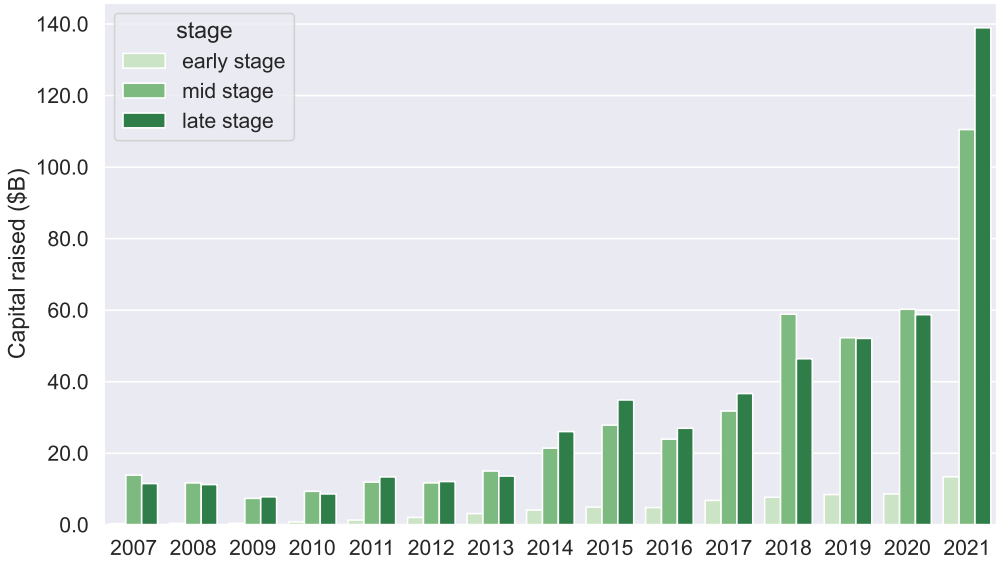

Fig. 3: Capital raised per year for each stage

imbalance is in favor of early stage rounds, when looking at the corresponding capital raised for each type, late stage rounds are over-represented, due to their nature, as illustrated in Figure 3

3.1.2. Market data

We fit the model on various public equity benchmarks, as well as on risk-free assets (treasury bills). To get the data, we use two sources: Yahoo Finance [2] and the Federal Reserve Bank of St. Louis (FRED). We access both websites through their APIs. For the risk-free asset, we use the 3-month treasury bill rate [3], which is convenient for this analysis as our baseline model uses a time grid of three months to fit the model, and is the standard in the existing literature.

We develop a Python script that allows to fetch the corresponding data and estimate the model automatically. Given a dataset of funding rounds, the algorithm automatically infers the correct dates for the time series and fetches the APIs to retrieve the benchmarks and T-bills data. The only required parameter is the ticker of the asset (typically `TB3MS` for the 3-month treasury bill rate, `^SP500TR` for the S&P500 Total Return Index, `NDAQ` for the

[2] Yahoo Finance, `https://finance.yahoo.com/lookup/`

[3] Board of Governors of the Federal Reserve System (US), 3-Month Treasury Bill Secondary Market Rate, Discount Basis [TB3MS], retrieved from FRED, Federal Reserve Bank of St. Louis; `https://fred.stlouisfed.org/series/TB3MS`

NASDAQ, `^RUT` for the Russell 2000).

3.2. Methodology

We use the same methodology as Cochrane (2005). The heavily skewed distribution of returns imposes the use of a log-returns to model the equity value of the firm:

$$\begin{cases} \mathrm{d}\ln \mathrm{V} = (\mathrm{r}^{\mathrm{f}} + \gamma)\mathrm{dt} + \delta(\mathrm{d}\ln \mathrm{V}^{\mathrm{m}} - \mathrm{r}^{\mathrm{f}}\mathrm{dt}) + \sigma \mathrm{dB} \\ \mathrm{d}\ln \mathrm{V}^{\mathrm{m}} = \mu_{\mathrm{m}}\mathrm{dt} + \sigma_{\mathrm{m}}\mathrm{dB}^{\mathrm{m}} \end{cases} \tag{1}$$

Where dB is a standard Brownian motion, V is the value of the asset (firm equity), r^f is the risk free rate, γ the intercept, δ the slope, V^m is the value of the market and σ is the volatility of the value process. In line with previous works, we assume $[B, B^m]_t = 0$. In discrete time (for a time step $\Delta t = 1$) the model is:[4]

$$\ln\left(\frac{V_{t+1}}{V_t}\right) = \ln R^f_{t+1} + \gamma + \delta(\ln R^m_{t+1} - \ln R^f_{t+1}) + \epsilon_{t+1} \tag{2}$$

Where $\epsilon_{t+1} \sim \mathcal{N}(0, \sigma^2\Delta t)$ follows a normal distribution, and

$$R^f_{t+1} = 1 + r^f_{t+1}, \quad \text{and} \quad R^m_{t+1} = 1 + \frac{V^m_{t+1} - V^m_t}{V^m_t}$$

With these assumptions, the value of the firm V_{t+1} follows a log-normal distribution with parameters:

$$\begin{aligned} \mu_{t+1} &= \ln R^f_{t+1} + \gamma + \delta(\ln R^m_{t+1} - \ln R^f_{t+1}) \\ \sigma^2_{t+1} &= \sigma^2 \end{aligned}$$

Hence the probability density function of V is:

$$P(V_{t+1}|\mathcal{F}_t) = \frac{1}{\sqrt{2\pi}\sigma_{t+1}V_t} \exp\left(\frac{(\ln(V_t) - \mu_{t+1})^2}{2\sigma^2_{t+1}}\right) \tag{3}$$

[4]The details of the derivation are available in the appendix of the paper and in Cochrane's original appendix available online at `https://www.johnhcochrane.com/s/venture_capital_appendix.pdf`

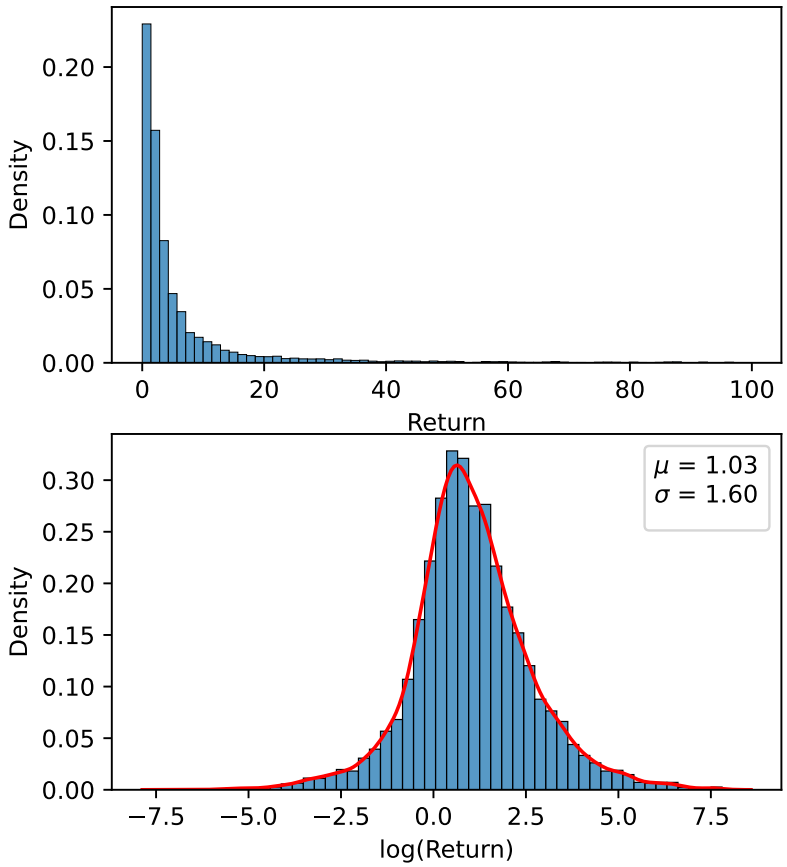

Fig. 4: Returns and log-returns from enhanced Crunchbase data

In the simulation, we set up a value grid $\Omega = [V_{\min} = V_0, V_1, \ldots, V_N = V_{\max}]$.

$$P(V_{t+1} \in [V_i, V_{i+1}[\,|\,\mathcal{F}_t) = \int_{V_i}^{V_{i+1}} P(V\,|\,\mathcal{F}_t)\mathrm{dV}, \tag{4}$$

with $\{V_{\min}, V_{\max}\}$ large enough such that

$$\sum_{i=1}^{N} P(V_{t+1} \in [V_i, V_{i+1}[\,|\,\mathcal{F}_t) \approx 1$$

Given the value distribution $P(V_t)$ at the beginning of period, we compute the joint probabilities of each different outcomes of the project at the end of the period:

1. Successful exit, either IPO or acquisition
2. Operating, the firm remains private
3. Bankrupt, the firm closes

Each of these three outcomes is characterized by an "exit flag" in the model, used to branch

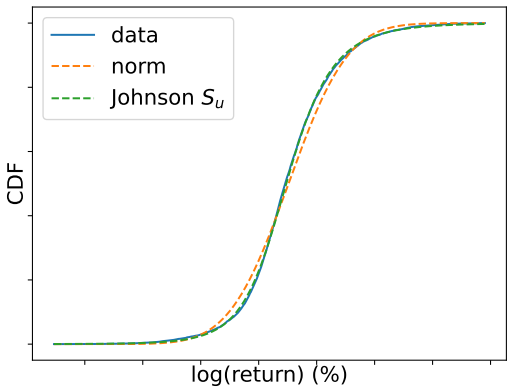

Fig. 5: log-returns CDF follows a normal distribution

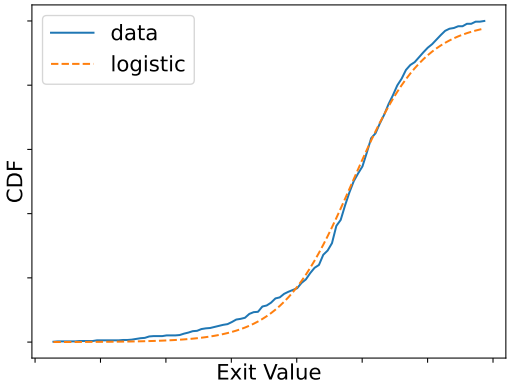

Fig. 6: Exit value CDF follows a logistic distribution

out into each probability calculation. The firm can get new financing, with probability:

$$P(\text{Exit}, V_t) = P(V_t)P(\text{Exit} \mid V_t) \tag{5}$$

The probability of getting an exit given a certain value V_t is modelled by a logistic function in value:

$$P(\text{Exit} \mid V_t) = \frac{1}{1 + e^{-a(\ln(V_t) - b)}}$$

The firm can fail and close at the end of the period, with probability:

$$P(\text{Close}, V_t) = P(V_t)P(\text{Close} \mid V_t)[1 - P(\text{Exit} \mid V_t)] \tag{6}$$

Where the probability of going bankrupt is modeled by a linearly decreasing function in value. The parameter k acts as a threshold for the firm value, above which we consider the probability of the firm going bankrupt to be zero:

$$P(\text{Close} \mid V_t) = \left[1 - \frac{V - V_0}{k - V_0}\right] \mathbb{1}_{V \leq k}$$

Finally, the remaining firms are those still private at the end of the sample, with probability:

$$P(\text{Private} \mid V_t) = P(V_t)[1 - P(\text{Exit} \mid V_t)][1 - P(\text{Close} \mid V_t)] \tag{7}$$

Using the above definitions, we compute the value distribution of the firms that remain private (the value of the others becomes irrelevant, as they exited) in the next period, and

continue the process until the end of the sampling period:

$$P(V_{t+1}) = \sum_{V_t} P(V_{t+1}|V_t)P(\text{Private}, V_t)$$

Where $P(V_{t+1}|V_t)$ is given in Eq. (1).

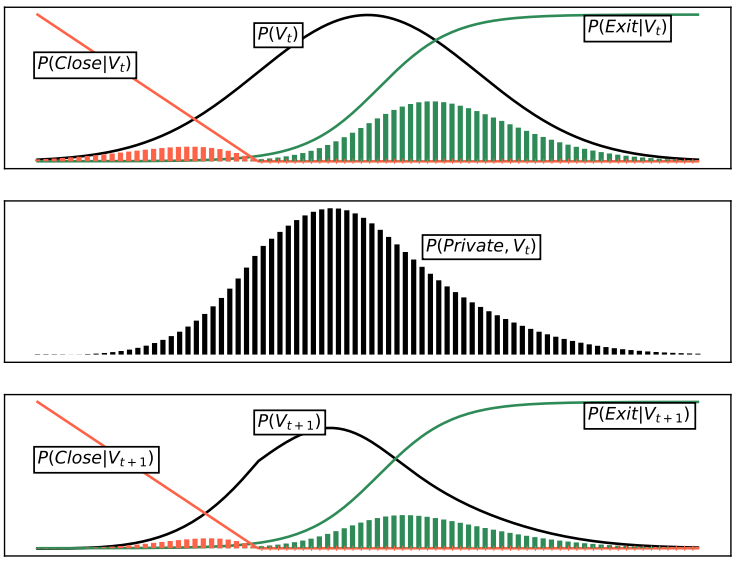

Fig. 7: Illustration of the simulation process

For each event, we decompose the probability functions further, as we do not necessarily have good data for every observation (amount raised or return may be missing or dates can be wrong). For that matter, we introduce three parameters: d is the fraction of all rounds with correct data (neither return nor date missing). We also assume a uniformly distributed measurement error probability in the data-set, denoted by π. With probability $1-\pi$ the data records the true value. With probability π the data erroneously records a value uniformly distributed over the value value grid Ω. Finally, c represents the fraction of all out-of-business rounds with good date. For a given data point x, with value V_t^x at t, depending on the project state (exited, operating, or closed) and observed data fields, it falls into one of the following categories:

- Category 1, observations of exits with good return and good date: the probability of observing a data point with good data and good date is d times the probability of a new round at age t with value V_t^x. As previously stated, we assume that a fraction

π of the data is erroneous. An erroneous observation is uniformly distributed over Ω. Accounting for data error, the resulting probability of seeing a data point with good dates and return is:

$$P_1(V_t^x) = d(1-\pi)P(\text{Exit}, V_t^x) + \frac{d\pi}{|\Omega|}\sum_{V_t \in \Omega} P(\text{Exit}, V_t)$$

- Category 2, observations of exits with bad return but good date. This records rounds for which we know some financing events happened for the given value, although we do not have the return associated with this event. This corresponds to the probability of having a new financing (conditional on having bad return data):

$$P_2(V_t^x) = (1-d)\sum_{V_t \in \Omega} P(\text{Exit}, V_t)$$

- Category 3, observations of firms remaining private at the end of the period:

$$P_3(V_t^x) = \sum_{V_t \in \Omega} P(\text{Private}, V_t)$$

- Category 4, observations from closed firms, with good dates. The probability of observing a firm closing at the end of sample if simply the integral of the probability density for this event at time t. We also account for the fraction of observations in the dataset that have bad dates:

$$P_4(V_t^x) = c\sum_{V_t \in \Omega} P(\text{Close}, V_t)$$

- Category 5, observations from closed firms, with bad dates. There is a fraction $1-c$ of the records that have bad dates:

$$P_5(V_t^x) = (1-c)\sum_{\tau=\tau_0}^{\tau_N}\sum_{V_\tau \in \Omega} P(\text{Close}, V_\tau)$$

The complete probability distribution is thus:

$$f(x) = \sum_{i=1}^{5} P_i(V_t^x)\mathbb{1}_{x \in C_i} \tag{8}$$

Next, we compute the likelihood of the parameters given any observation x. The likeli-

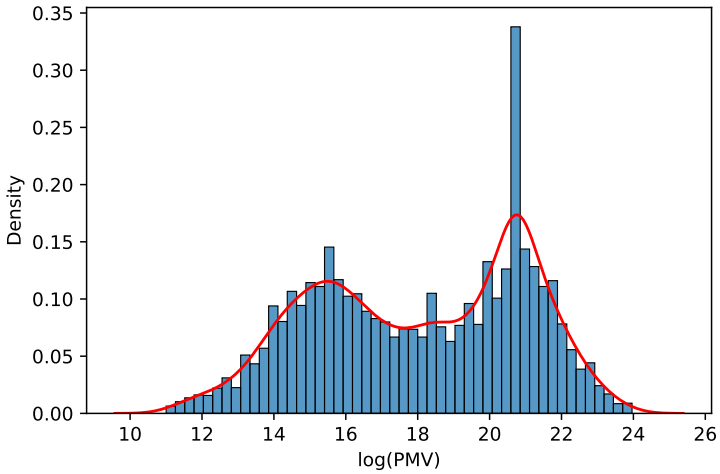

Fig. 8: Distribution of the logarithm of post-money valuations

hood function takes the form, with the sample $\mathbf{x} = (x_1, x_2, \ldots, x_n)$:

$$\mathcal{L}(\theta \,|\, \mathbf{x}) = f(\mathbf{x} \,|\, \theta), \quad \theta = (\gamma, \delta, \sigma, k, a, b, \pi)$$

where γ is the intercept of the log-model, δ the slope, σ the volatility of the return process, k is the bankruptcy threshold, a and b are the parameters for the exit probability function and π is the parameter controlling for data errors (proportion of erroneous data points). We solve for the following problem:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta \,; \mathbf{x}) \tag{9}$$

3.2.1. Accounting for missing data

Although Crunchbase stands as one of the current top standards for venture capital data, it suffers from the same drawbacks as most previous databases: data scarcity. One of the advantages of Crunchbase is the large number of recorded companies and funding rounds. However, this study requires more than information about whether or not a new financing round happened. To compute and handle returns we need several attributes, the date of the funding round, the funded company, the amount of money raised, and the post-money valuation (as well as previous funding rounds data to account for dilution). To our knowledge, the number of rounds recorded is larger than in any other databases (see Section 3), but a substantial amount of PMVs is missing (and more rarely, the amount of money raised). To overcome this problem, we develop an estimation model based on statistical tools.

Estimating post-money valuations is a regression task. Given a set of attributes, we infer what should be the value of the firm. We account for the fact that the post-money valuation, an accounting value given by analysts, is a biased estimate of the true value of the firm. This phenomenon is clearly illustrated in the original dataset, in which we observe an abnormal spike at the billion values (see Figure 8). This is a well known human bias so-called "round number bias" (see, *e.g.*, Hervé and Schwienbacher, 2018). For these reasons, we are not interested in the precise value of the firm after a financing event, but rather in an unbiased estimate of the order of magnitude of the business value. Note that post-money valuations follows a bi-modal distribution, centered roughly around the 2-3 million value ($e^{15} \approx 3e6$) and the billion value ($e^{21} \approx 1e9$).

Symbol	Description
T	Date of the round (in days, relative to 1/1/1926)
ΔT	Number of days since last financing event
M	Amount of money raised ($)
ΔM	Difference of money raised since last round ($)
N	Number of investors for the current round
R	Lead investor rank for the current round
S	Industry sector (categorical)
G	Geographical position (categorical)

Table 1: Features used for post-money valuation regression

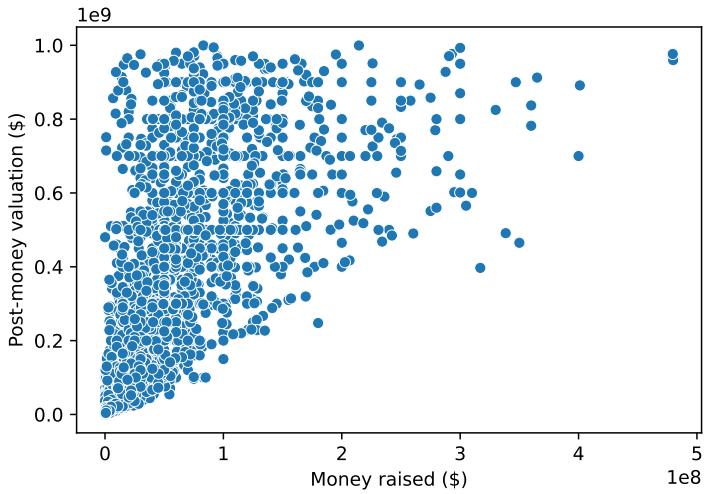

Fig. 9: Post-money valuation compared to money raised (for firms valued less than a billion)

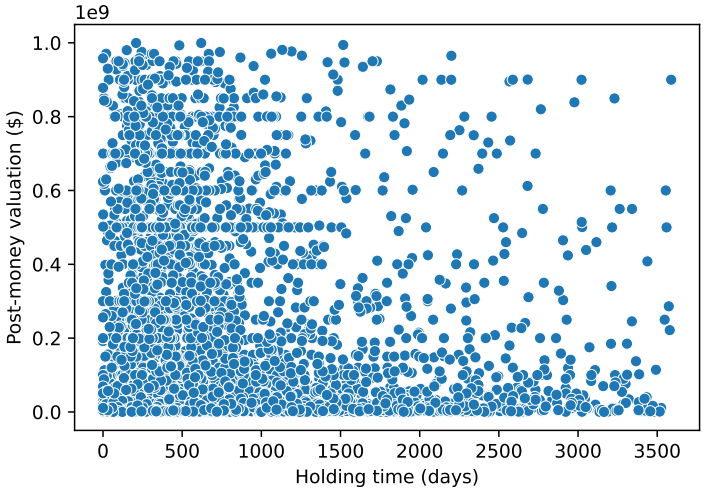

Fig. 10: Post-money valuation versus holding time.
For companies valued less than a billion.

When considering what could contribute to the value of a firm after a financing event, we select features based on their relevance and availability. The most important feature is the amount of money raised M_i for round i. Not surprisingly, the amount raised is highly correlated to valuation: the more money raised, the bigger the valuation (see Figure 9). This feature is widely available across the dataset, and thus removing the data points where it is not available should not introduce a significant bias. We also consider the difference in money raised between the current and previous rounds: $\Delta M_i = M_i - M_{i-1}$ (and $\Delta M_0 = 0$). This variable is a good indicator of the state of the firm and the confidence from the investors. If the amount raised since last round is booming, this would indicate that the product has a lot of success, that investors believe in the company, and thus, corresponds to a higher valuation. On the contrary, if the firm's business is slowing down along with less willing investors, this should translate to a lower valuation. The next set of features is temporal attributes: the date of the round T_i and the amount of time since last financing event $\Delta T_i = T_i - T_{i-1}$ (and ΔT_0 is the number of days between the firm creation and the first round of financing). The date is expressed in days, with 01/01/1926 as the arbitrary starting date. The date of the round carry information about the time-varying economic context that influences valuations (see, for example, the latest boom in valuations during the last three years, or during the internet bubble). The time between two rounds is an indicator of the dynamics in the firm's life cycle.

More frequent rounds often induce larger valuations (see Figure 10). Information on investors is also useful to predict a firm valuation. It has been well documented that network plays a central role in the venture capital industry (see Alexy, Block, Sandner, and Ter Wal,

Country	Avg. PMV	Sector	Avg. PMV
U.K.	\$348M	Energy	\$238M
France	\$470M	IT Hardware	\$321M
Germany	\$670M	HC Services	\$498M
USA	\$753M	Other	\$656M
Sweden	\$769M	Media	\$758M
Japan	\$801M	Consumer Goods	\$799M
New-Zealand	\$974M	Transportation	\$807M
India	\$1,470M	Software	\$810M
Korea	\$2,637M	Commercial Products	\$942M
China	\$3,752M	Pharma & Biotech	\$1,119M

Table 2: Average PMV for the top 10 countries and sectors on Crunchbase

2012). Given this fact, we use two investor-related features. First, the number of investors participating in a given round. Again, given the importance of networking in VC, a larger number of investors participating in a round indicate a potentially (or expected to be) successful firm. Second, a feature directly extracted from Crunchbase provides, for each round, who is the leading investor. Crunchbase maintains, for each investor and organization, a rank, measuring how active and influential they currently are. Given these two fields, we record the rank of the lead investor in each round. Finally, we include two categorical variables: the industry sector and the geographical location. Fundings and valuations largely differ from one industry to the next (see Table 2) and thus seems to be a relevant predictor.

We use the python package Sklearn (Pedregosa, Varoquaux, Gramfort, Michel, Thirion, Grisel, Blondel, Prettenhofer, Weiss, Dubourg, Vanderplas, Passos, Cournapeau, Brucher, Perrot, and Duchesnay, 2011) to build and train the models. For this regression task, we test several machine learning models, including Support Vector Machines (SVM) regressor, decision trees, and ensemble methods (AdaBoost, Gradient Tree Boosting). These models have proven their efficiency in complex regression problems and can easily fit non linear distributions in high dimensional environments. For such complex problems, ensemble methods are often the best performing architectures. To confirm this intuition, we use an automated procedure to find the best architecture among an exhaustive list of models. This procedure known as Auto-ML (for Automated Machine Learning), leverages recent advantages in Bayesian optimization, meta-learning, and ensemble construction to find the appropriate steps and parameters for the usual machine learning workflow:

- Preprocessing and data cleaning
- Features construction and selection
- Model family selection

- Model hyperparameters optimization

The particular implementation used in this project is "AutoSklearn" Feurer, Eggensperger, Falkner, Lindauer, and Hutter (2021). As this technology is still in its early development, we do not rely on it for fine tuning, but only to get a global picture of the best and worst architectures.

Finally, the returns are computed from round to IPO, taking into account the dilution. We first compute the value of equity at the final round (IPO or acquisition) for an investor who entered at each round i:

$$x_i = \underbrace{\frac{m_i}{v_i}}_{\substack{\text{initial stake} \\ \text{of investors i}}} \times \underbrace{\frac{v_i - m_{i+1}}{v_{i+1}}}_{\substack{\text{proportion of old} \\ \text{equity at round } i+1}} \times \ldots \times \underbrace{\frac{v_{n-1} - m_n}{v_n}}_{\substack{\text{proportion of old} \\ \text{equity at round } n}}$$

Where $m_i, m_{i+1}, \ldots, m_n$ are the amount raised from investors at each round, $v_i, v_{i+1}, \ldots, v_n$ the equity value at each round (including IPO) and x_i is the percentage of equity owned by investors at the exit event. The return for these investors is then,

$$R_i = \frac{\overbrace{v_n \times x_i}^{\substack{\text{value owned} \\ \text{at exit}}} - m_i}{m_i}$$

For example, suppose firm XYZ raises \$1M during its seed round and is then valued \$2M. Investors possess 50% of the equity. In the next round, the firm raises \$5M with a post money valuation of \$15M. Previous investors hold 50% of the \$10M left that does not belong to new investors, that is \$5M in equity, or 33% of the firm value. If the firm was to exit at this round, they would get a return of $(0.33 \times 15 - 1) \div 1 = 4$. That is a $\times 4$ return (400%). In case of bankruptcy, we consider that investors lose all their initial stake, giving a return of -1.

4. Results

4.1. Post-money valuation regression

This section presents the results for the regression models of post-money valuation. The performance of the models has a high impact on the future performance of the maximum likelihood estimation, as the output is directly used to compute returns. The first consideration was the choice of the model. As our problem is highly non-linear, simple models like linear, Ridge and Lasso, or logistic regressions, perform very poorly. The best performing models

Model id	Rank	Weight	Type	Cost
12	1	0.1	HistGradientBoosting	0.5475
15	2	0.1	HistGradientBoosting	0.5812
16	3	0.04	KNeighborsRegressor	0.7581
19	4	0.02	LinearSVR	1.0513
2	5	0.22	RandomForestRegressor	0.5288
32	6	0.02	KNeighborsRegressor	0.6689
35	7	0.02	ARDRegression	0.5774
42	8	0.04	KNeighborsRegressor	0.8135
46	9	0.02	HistGradientBoosting	0.5735
48	10	0.02	HistGradientBoosting	0.6192
54	11	0.02	HistGradientBoosting	0.5561
61	12	0.02	HistGradientBoosting	0.5859
68	13	0.22	HistGradientBoosting	0.5477
77	14	0.12	DecisionTreeRegressor	0.5843
81	15	0.02	RandomForestRegressor	0.5280

Table 3: Output of Auto-Sklearn procedure
Column "Rank" corresponds to the relative rank of a given model based on its score. "Cost" corresponds to the loss of the model on the validation set. "Weight" corresponds to to the weight attributed to a particular model for the final prediction.

according to this process is, as expected, ensemble methods and especially boosting trees. The specific model highlighted by the procedure is the Histogram-based Gradient Boosting Regression Tree. This architecture is particularly useful when dealing with large datasets (more than 100,000 observations) for its low memory footprint and small computation time. The main drawback of this model is that it acts as a black-box, unlike other models such as the classical Gradient Boosting Regressor.

Because of the very high variance in both post-money valuations and funding amounts, the dataset has been further divided into two subsets, roughly around the median (median $\simeq$ 18.5). This split between models stems from the observation that funding dynamics differ from one stage to another. We summarize the results of the different models in Table 4. The best performing model overall is the baseline logarithmic model. When looking at regression on log-values, it clearly outperforms every other model, both in terms of R^2 and MAE. Only the "Big" model beats the MAPE score by only 0.59%. When looking at absolute values, there is no clear cut. "Small" model is the best in terms of R^2, but worse than the baseline-ln and "Big" models in terms of MAPE. When looking at absolute values, the median absolute error is not relevant for "Big" and "Small". Indeed, when looking at smaller firms, their predicted valuation is also scaled down, and thus the median error for the "small" model is

	Logarithmic			Absolute		
Model	R^2	**MAE**	**MAPE**	R^2	**MAE**	**MAPE**
Baseline, abs	0.65	7.78e-1	7.92%	0.57	4.63e+7	1057.20%
Baseline, ln	0.94	3.67e-1	2.91%	0.56	2.49e+7	57.03%
Small	0.78	3.71e-1	3.57%	0.70	1.65e+6	64.92%
Big	0.70	3.75e-1	2.32%	0.43	3.12e+8	55.36%
Ensemble	0.95	3.77e-1	2.97%	0.54	2.23e+7	62.38%

Table 4: Model performance for PMV regression
MAE = Median absolute error, MAPE = Mean absolute percentage error. "Baseline, abs" contains all variables and is optimized to fit absolute PMV values. "Baseline, ln" contains all variables and is optimized to fit logarithmic PMV values. "Small" and "Big" refer to two models fit on two subset of the data, based on the bimodal distribution. "Ensemble" is the output of the Auto-ML procedure using an ensemble of models to make a prediction (see Table 3)

structurally smaller than that of the "big" model. The baseline-ln model beats by a factor of two the baseline-abs model, although the order of magnitude is similar. MAPE for the baseline-abs model is completely off, confirming its poor regressive power. The baseline-ln model acts as a good trade-off between each statistics overall.

Figure 11 shows that log-values are accurately predicted but any small error in the right tail leads to significant error on the absolute firm value. Median absolute error is robust to such extreme values. Mean absolute percentage error on the contrary is sensitive to errors in the left tail, as firms with a very small valuation yield high MAPE if the predicted valuation is even one order of magnitude higher than the ground truth.

The overall performance is good, but deteriorates quickly as we go to higher valuations, due to the log-transform. However, data of the right hand is very sparse, and thus should not contribute a lot to the final predictions. This matter will be investigated in the next section, to see which observations drive the estimates the most. Also, since we use a log-normal model (Eq. 1), the values that are ultimately used for the calculations are the log-PMVs. Finally, the mean absolute percentage error is driven by outliers. Table 5 show descriptive statistics for the MAPE. Only 219 observation have a MAPE greater than 3σ ($\sigma = 97.25\%$). Filtering these outliers yields very decent numbers, as low as 41% average error. The distribution of outliers does not follow a particular pattern. All small, medium, and large firms occasionally suffer from large regression errors in both directions.

As previously stated, one of the biggest disadvantages of the histogram-based gradient boosting regressor is its "black-box" design, and there is no real way of interpreting the model. To get an idea of the feature importance, we run a mock (classical) gradient boosting model (GBR). The main difference is that the base GBR uses the continuous values in the dataset whereas histogram-based GBR bins these values (hence the term "histogram") and works

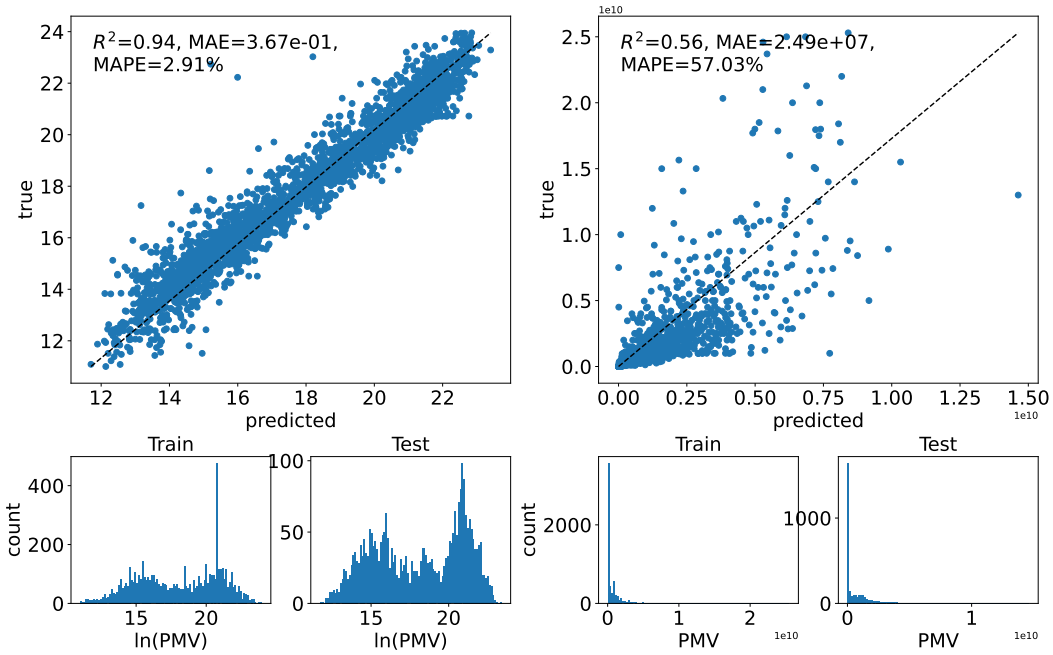

Fig. 11: Visualization of the test results for the best performing model

Stat	**MAPE**	$\|.\| < 3\sigma$
N	2750	2531
μ	62.97	41.80
σ	97.25	31.71
min	0.01	0.01
25%	17.66	16.46
50%	39.71	35.61
75%	69.17	60.57
max	1324.36	147.40

Table 5: Test set MAPE analysis for full model

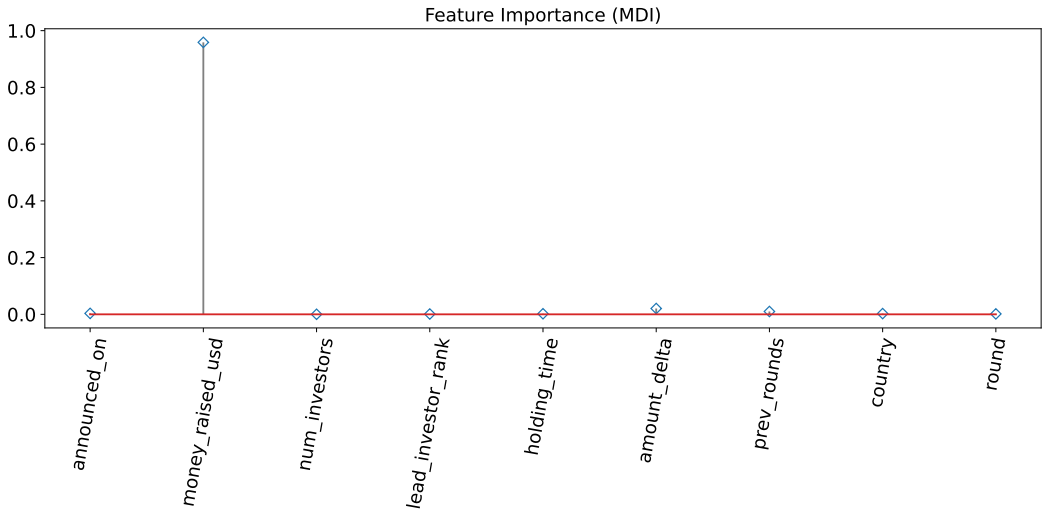

Fig. 12: Feature importance for the gradient boosting model
log results: $R^2 = 0.93$, $MAE = 4.19e-1$, $MAPE = 3.23\%$
abs results: $R^2 = 0.56$, $MAE = 3.08e+7$, $MAPE = 70.05\%$

on the ordinal values of each bins. Numbers are very close to those of the histogram-based model, although still worse (by almost 20% for MAPE on absolute values). This poorer performance may be due to overfitting, since the non histogram-based version takes into account each feature individually rather than bins.

Figure 13 show the train and test loss across boosting iterations (*i.e.* number of estimator trees). The model reaches overfitting quickly at around 30 estimators, as the the training loss keeps decreasing but the test loss start increasing. Figure 12 depicts the overfitting, as the most important feature for determining the post money valuation is the amount of money raised. This result was expected but the more surprising is how little information other variables carry according to the model. Every other feature has practically zero prediction power (except "amount delta" and "prev rounds").

4.2. Global Market

In this section we present and analyze results for all the observations, or the "baseline" model. These are the results for the enhanced dataset, with returns from round to exit (not round-to-round returns, which are not discussed in this project). Table 6 presents descriptive statistics of the completed data set using the machine learning approach. We observe one of the co-effects of the selection bias, as the number of rounds available increases, the quality of the data increases (parameter d measures the proportion of good data, *i.e.* good return and good dates, where we need to observe post-money valuations and exit value). Note that more than 85% of the observations (or more than 100,000 entries) occurred at a date after

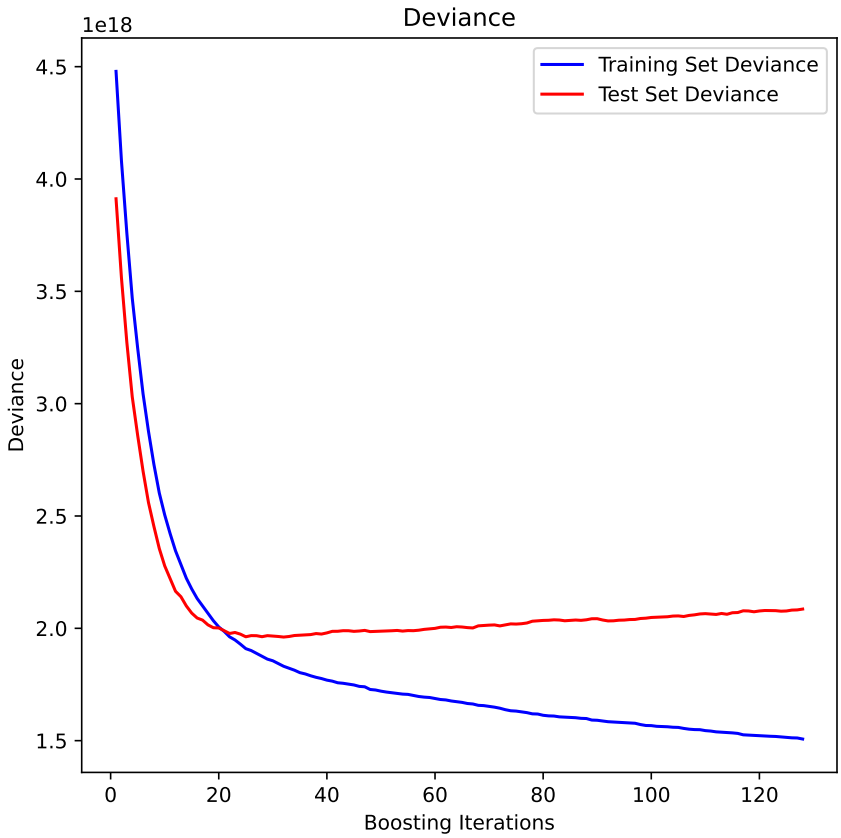

Fig. 13: Base gradient boosting train and test loss across boosting iteration

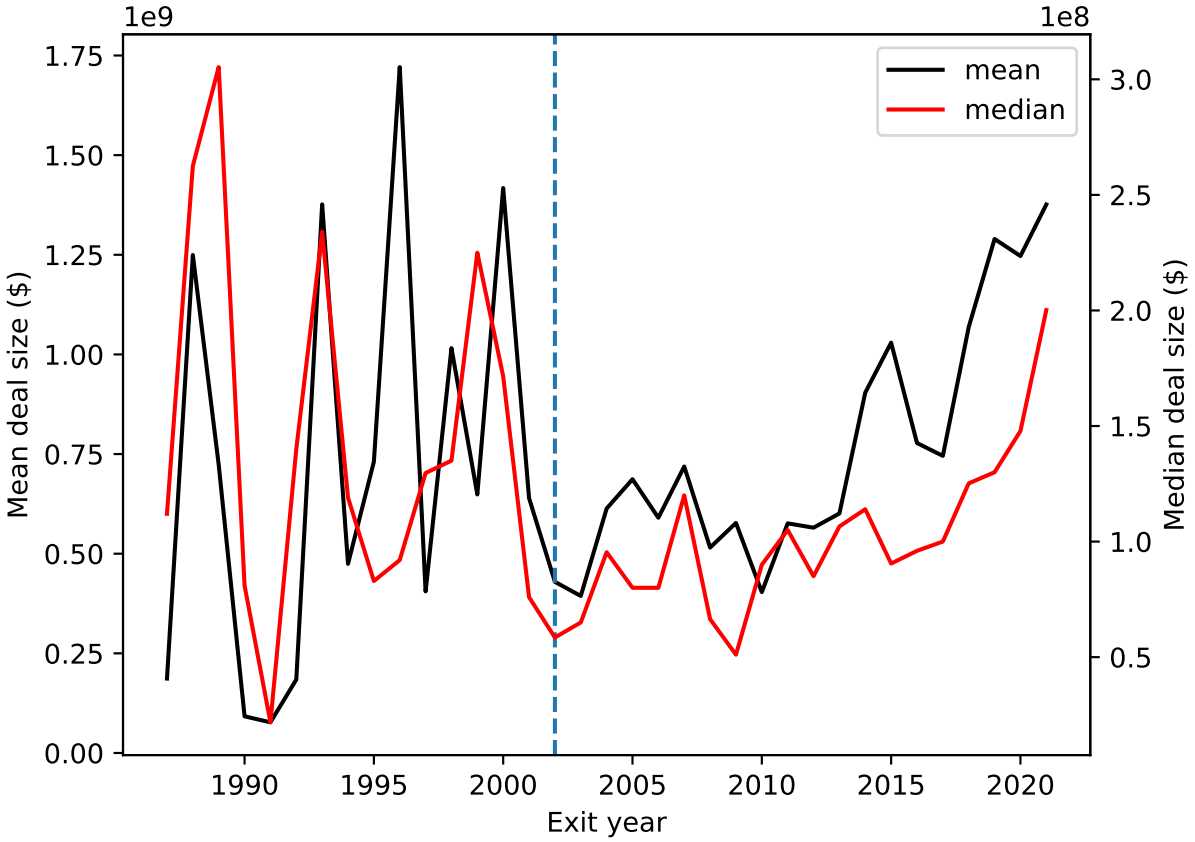

Fig. 14: Average and median deal size in dollars from 1987 to 2021 using Crunchbase data

January 2010. This leads to a heavy selection bias for observations prior to this date, as only the "biggest" transactions are likely to be recorded. Indeed, the mean and median deal size across years show a positive trend from 2002 onward, whereas observations for years prior to 2002 is erratic, without a clear trend, and are on average even higher than for the recent period (Figure 14). This bias likely drives estimates because of the high impact of the pre-2010 observations. For comparison, Cochrane (2005) has almost 20,000 observations for the period 1987-2000. For this reason, we choose to carry the analysis on data starting from 2010. This includes IPOs, acquisitions, closings, and funding rounds that took place starting from January 2010. The proportion d of observations with good data remains almost identical.

Table 8 reports results for the maximum likelihood estimation on the dataset containing observations from January 2010 to March 2022. Derivation for the estimates are obtained

		Rounds			
	All	**1**	**2**	**3**	**4+**
Total	118590	48973	29676	17934	22007
IPO	6168	1220	1266	1158	2524
Acq.	17833	62865	5112	3177	3259
Closed	3500	1667	1024	498	311
Private	91089	39801	22274	13101	15913
d	36%	23%	31%	37%	58%

		Industries			
	All	**Tech**	**Retail**	**Health**	**Other**
Total	118590	80474	19463	6448	12153
IPO	6168	4773	779	167	449
Acq.	17833	12321	2806	620	2081
Closed	3500	2229	556	117	596
Private	91089	61151	15322	5544	9027
d	36%	37%	36%	25%	32%

Table 6: Descriptive statistics for the enhanced data set
Includes data from January 1990 up until March 2022. Most of the observations occurred after 2010.

as follows. Recall Eq. (1). Taking the expectation and variance yields:

$$\mathbb{E}[\ln r] = \gamma + \mu_{\ln r_f} + \delta(\mu_{\ln r_M} - \mu_{\ln r_f}) \tag{10}$$

$$\mathbb{V}[\ln r] = \delta^2 \sigma^2_{\ln r_M} + \sigma^2 \tag{11}$$

Since the simulation is performed on a quarterly basis (3 month grid), we multiply by four to annualize results and by 100 to get percentages. To get the results for the arithmetic returns, we take the expectation and variance of a log-normal variable, with $R = r + 1$ and $\mu = \mathbb{E}[\ln R]$:

$$\mathbb{E}[R] = \exp\left(\mu + \frac{1}{2}\sigma^2\right) - 1 \tag{12}$$

$$\mathbb{V}[R] = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2) = (\exp(\sigma^2) - 1)(\mathbb{E}[R] + 1)^2 \tag{13}$$

For the period 2010-2022, we use the following values for calculations (retrieved from Yahoo Finance and the FRED, see 3.1.2):

- 3 months T-bill: $\mu_{\ln r_f} = 0.48\%$
- S&P500: $\mu_{\ln r_M} = 4.50\%$ and $\sigma_{\ln r_M} = 7.94\%$
- NASDAQ: $\mu_{\ln r_M} = 5.11\%$ and $\sigma_{\ln r_M} = 9.22\%$

S&P500

Industry	γ	δ	σ	k	a	b	π
Baseline	22.49	1.26	40.74	58.00	0.52	9.98	25.25
	(0.56)	(0.13)	(1.63)	(0.54)	(0.00)	(0.00)	(0.75)
Tech	23.35	1.65	46.47	52.76	0.54	9.99	22.43
	(0.57)	(0.10)	(2.41)	(1.12)	(0.00)	(0.00)	(1.12)
Retail	22.08	1.72	36.07	63.14	0.53	10.00	28.52
	(0.99)	(0.08)	(6.43)	(1.99)	(0.00)	(0.00)	(3.20)
Health	20.53	0.93	23.82	85.99	0.55	10.00	54.83
	(1.59)	(0.21)	(17.89)	(4.10)	(0.01)	(0.00)	(5.78)
Other	20.17	0.60	48.34	39.85	0.49	10.00	14.34
	(0.41)	(0.01)	(2.30)	(0.80)	(0.00)	(0.00)	(1.01)

NASDAQ

Industry	γ	δ	σ	k	a	b	π
Baseline	23.00	1.17	40.70	58.11	0.52	9.99	25.24
	(0.86)	(0.01)	(0.79)	(0.57)	(0.00)	(0.00)	(0.79)
Tech	24.92	1.27	39.18	63.76	0.53	10.00	28.73
	(0.36)	(0.06)	(1.81)	(0.57)	(0.00)	(0.00)	(1.20)
Retail	23.69	1.41	35.85	64.15	0.53	9.98	28.56
	(1.41)	(0.23)	(8.41)	(3.33)	(0.00)	(0.00)	(4.08)
Health	21.23	0.77	23.63	86.44	0.54	10.00	55.00
	(1.44)	(0.20)	(15.58)	(3.50)	(0.01)	(0.00)	(5.44)
Other	19.65	0.73	48.36	39.89	0.49	10.00	14.37
	(0.75)	(0.15)	(2.59)	(1.09)	(0.00)	(0.00)	(1.41)

RUSSELL 2000

Industry	γ	δ	σ	k	a	b	π
Baseline	23.48	1.03	40.95	56.87	0.52	10.00	24.61
	(0.23)	(0.04)	(1.21)	(0.29)	(0.00)	(0.00)	(0.81)
Tech	26.03	1.10	38.78	64.72	0.54	9.97	28.96
	(0.38)	(0.06)	(2.00)	(0.72)	(0.00)	(0.00)	(1.18)
Retail	24.32	1.27	37.98	59.76	0.53	10.00	27.45
	(1.17)	(0.16)	(6.88)	(2.72)	(0.00)	(0.00)	(3.60)
Health	21.96	0.70	26.24	82.62	0.55	10.00	52.41
	(1.26)	(0.16)	(7.13)	(0.92)	(0.01)	(0.00)	(6.03)
Other	20.28	0.59	48.35	39.88	0.49	10.00	14.40
	(0.49)	(0.08)	(1.92)	(0.49)	(0.00)	(0.00)	(1.05)

Table 7: Estimated parameters by maximum likelihood, from 2010.
Values for γ, σ are given as annualized percentages, and values for k, π are given as percentages. Standard errors are in parenthesis.

S&P500

Industry	E[ln R]	E[R]	α	β
Baseline	28.04	38.60	32.14	1.37
	(41.96)	(46.52)	(0.67)	(0.10)
Tech	30.45	44.40	36.03	1.81
	(48.28)	(54.43)	-	-
Retail	29.46	38.65	30.05	1.86
	(38.56)	(42.68)	-	-
Health	24.74	28.84	24.07	0.98
	(24.93)	(26.83)	-	-
Other	23.05	36.41	33.13	0.64
	(48.58)	(53.79)	-	-

NASDAQ

Industry	E[ln R]	E[R]	α	β
Baseline	28.20	38.68	32.65	1.27
	(41.76)	(46.30)	(0.51)	(0.10)
Tech	30.52	40.64	34.11	1.39
	(40.46)	(45.03)	-	-
Retail	29.85	38.66	31.50	1.53
	(37.56)	(41.55)	-	-
Health	24.80	28.77	24.73	0.81
	(24.41)	(26.26)	-	-
Other	23.06	36.49	32.58	0.78
	(48.71)	(53.95)	-	-

RUSSELL 2000

Industry	E[ln R]	E[R]	α	β
Baseline	28.09	38.56	33.21	1.11
	(41.76)	(46.28)	(0.58)	(0.06)
Tech	30.92	40.77	35.08	1.19
	(39.75)	(44.24)	-	-
Retail	29.88	39.42	32.95	1.37
	(39.29)	(43.58)	-	-
Health	25.24	29.90	26.20	0.74
	(26.82)	(28.95)	-	-
Other	23.14	36.51	33.25	0.64
	(48.58)	(53.80)	-	-

Table 8: Implied estimates for $\mathbb{E}[\ln R]$, $\mathbb{E}[R]$, α and β

Implied estimates for the expected value and standard deviation for returns and log-returns, as well as α and β. Values given as annualized percentages (except for β). Standard deviations are in parenthesis.

- Russell 2000: $\mu_{\ln r_M} = 3.44\%$ and $\sigma_{\ln r_M} = 10.96\%$

The results for the baseline estimation show stable results across benchmarks. We obtain results for α and β using analytical formulas, and thus do not get the standard errors of these estimates.[5] We get these values using bootstrapping. Values for α are around 32% for all indices, and values for β vary between 1.11 for the Russell 2000 to 1.37 for the S&P500. Values above one for β suggest that venture capital is riskier (in terms of systematic risk) than any of these benchmarks, although not by much. On the contrary, we observe a very large α estimate for all benchmarks. We find a substantial arithmetic expected return for venture capital investments, with $\mathbb{E}[R] \simeq 38.60\%$ for each benchmark with a very high annual standard deviation (around 40%). This result is mostly driven by a high standard deviation (total risk) of venture capital with $\sigma \simeq 40\%$ in Eq. (12).

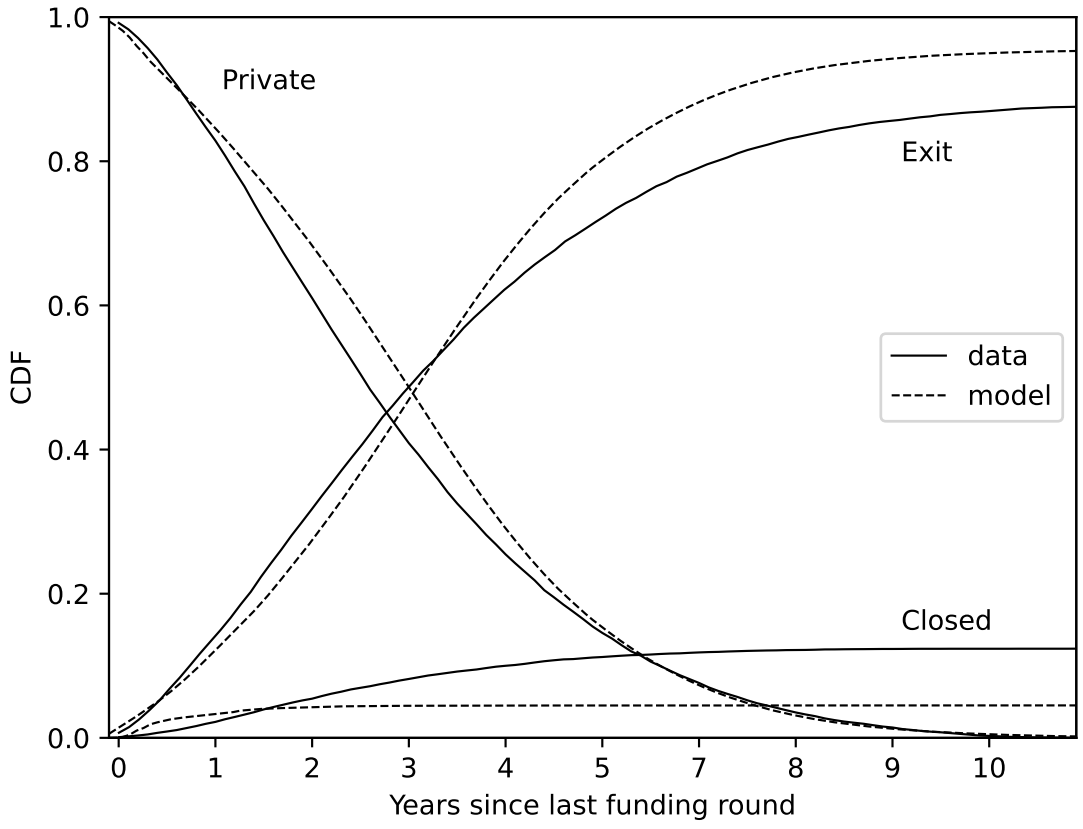

Fig. 15: Example simulation result for the optimal MLE parameters

Figure 15 shows the result of the maximum likelihood estimation as the cumulative distribution of the fates (exited, closed, or still private), as a function of the time since

[5]The risk and return of venture capital (Appendix), John H. Cochrane, 2005

last investment. Using estimated parameters from the MLE procedure, we run a simulation and plot the various probability distributions returned by the simulation, corresponding to equations in Section 3. We merge probabilities obtained for good and bad data. We observe an exponential decay for the firms remaining private, in line with Cochrane (2005). Nevertheless, our results point to a faster exit pace. After five years, more than 60% of the firms have gone public or have been acquired (50% in the original paper). Cochrane (2005) reports that a similar sharper decay in private firms is already present in the late part of his dataset. This trend has only accelerated since then, coherent with anecdotal evidence reported by the media and the industry. The model estimates quite well the true distributions, although the bankrupt probability distribution is consistently under-estimated (and thus the exit probability is over-estimated). This may be in part due to the simplistic assumption of the distribution (linear function of value below a threshold). In addition, the maximum likelihood estimation is not meant to perfectly replicate every single moment independently but rather the whole distribution. In that regard the estimation performs reasonably well.

The error measurement parameter π is estimated at 25%, that is, 25% of the measurements are treated as errors rather than extreme values in the logarithmic distribution. It does not discard extreme returns, but reasonable returns that occur in a very short time period, leading to very large annualized returns. Cochrane (2005) further argues that even after discarding true returns, venture capital is more about earning a large return over a few years rather than a relatively small return (¡2x) in a month. The cut-off value k measuring the maximum business value at which a firm can go bankrupt is estimated at 58% which is high but not incoherent. This means that above 58% of the initial value, the probability of a firm going bankrupt is 0. At 29% of its original value, this probability is 0.5. Finally, parameters for the exit probability function, a and b, are consistently estimated at 0.52 and 10.00. Getting constant results here should not be surprising, as these parameters only depends on the dataset and are independent of the choice of benchmark. Their direct interpretation suggests that there is a 50% probability of the firm going public during a given quarter from 1000% log-return. Moreover, this estimate is statistically significant at the 1% level, regardless of the initial choice of parameters (it always converges to the same values). Cochrane (2005) also finds a very high value of 380%.

The Refinitiv Venture Capital Index (TRVCI) seeks to replicate the return profile of the VC industry by constructing a theoretical dynamic portfolio in public, liquid assets that tracks the movements of the VC industry. We take this index as a comparison basis for our results. When running a regression analysis of this index from 2010, on respectively the S&P500, the NASDAQ and the Russell 2000, we find values for β of respectively 1.56, 1.50,

and 1.03, consistent with our aforementioned results. The S&P 500 yields the largest β and the Russell 2000 yields the lowest, close to one. However, our estimates show large positive α whereas the TRVCI exhibits an almost zero (positive) α.

4.3. Industry specific results

The individual estimations for each industry segment separately also yields consistent parameters across all benchmarks. All industries exhibit large positive intercepts γ for log-returns at around 20% (annualized). tech and retail industries are significantly riskier than the market with higher slopes δ, in particular compared to the S&P 500 ($\delta_{\text{Tech}} = 1.65$ and $\delta_{\text{retail}} = 1.72$). The total risk is still high but the health industry has the lowest deviation by far with $\sigma_{\text{health}} \simeq 25\%$, almost half of that of the "Other" group. This is consistent with a threshold value k that is also much higher at 82%. This indicates that health firms, even in later stages can still close with high probability. In contrast, in other industries, the estimate is closer to that of the global PE market. The lower variance and a higher value for k are also driven by the larger number of late IPOs or acquisitions the health industry together with the fact that the failure rate is relatively stable across industries. The measurement error parameter π is also higher for the health industry, indicating frequent high returns in small amount of time that the model interprets as outliers.

The exit probability distribution does not change across industries. Large volatilities yield the same high arithmetic returns, quite similar to the one observed in the base case. For the aforementioned reasons, the health industry has the lowest expected return at 30% and lowest α. Other industries are comparable to one another in terms of α. The β of the "Other" industry is the lowest ($0.64 \leq \beta \leq 0.78$), closely followed by the health industry ($0.74 \leq \beta \leq 0.98$). Retail industry exhibits a consistently higher estimate for β between 1.37 and 1.86, as well as the largest α above 34%.

Every industry sample contains several dozen thousands observations, except for the health industry, such that each sample should be representative of the corresponding market sector. The general takeaway is that health industry is much riskier than retail, tech, or the rest of the private equity market, but also carries a higher α. Firms in health industry overall exit faster, yielding high annualized returns for exiting firms, but firms also fail faster.

4.4. Security industry

We analyze the results specifically for the "security industry". The security business and especially cyber-security firms are of the uttermost importance in our connected world. These firms develop and implement new solutions to increase security in IT systems, protect

virtual assets, custom sensitive information, secure transactions and communications, and much more. From a market point of view, the cyber-security risk is real and has been priced (Florackis, Louca, Michaely, and Weber (2022) for example). It can impact markets as any other critical operational risk, and even be considered as systemic (inter-state cyber-attacks). We analyze this industry, as it is likely that firms behave differently than health firms or even tech firms as whole. Being able to detect the current trends in the venture capital market and asses risk for these firms is important to get a broad view of the market for any investor or actor who plans on engaging with this industry. We proceed as follows: Crunchbase maintains a list of attributes, such as categories and sectors, for each firm. These attributes vary depending on the business of the firm. Sectors correspond to a broad market classification (such as health or tech) and categories are more precise attributes. Firms in the security industry must belong to at least one of the following categories:

- Information Technology
- Network Security
- Cyber Security
- Security
- IT Infrastructure
- National Security
- Privacy
- Cloud Security
- Homeland Security
- Fraud Detection
- Spam Filtering
- Intrusion Detection

Results for companies whose business is related to one of these topics are shown in Tables 9 and 10. Once again, results are robust to the choice of the benchmark. We obtain a large arithmetic α, which is on par with the tech industry α, but still higher than most of the other industries. This higher alpha does not come with a higher β. In fact, the security industry β seems lower than that of the tech industry. It is still higher than the β of the global market, for each benchmark. Finally, the expected arithmetic returns are also the highest, above 40% annualized returns. These results tend to show that the cyber-security industry does not have a particularly low risk, but rather a lower systematic risk than the tech industry. A lower beta may be due to the counter-cyclical nature of the business: it reacts to negative events impacting other companies. This matter should be investigated further. These results confirm two observable trends. First, the cyber-security industry is one of the most valuable industry in the market, probably due to the increase demand for cyber-risk mitigation. Second, although more attractive than other sectors in the past years, it still behaves mostly as the rest of the market. The cyber-security industry is not a particular outlier, and does not exhibit very specific metrics that could indicate a very unique behavior for this sector.

S&P500

	γ	δ	σ	k	a	b	π
Baseline	25.61	1.43	41.81	58.36	0.53	10.00	21.14
	(0.85)	(0.17)	(3.93)	(1.76)	(0.00)	(0.00)	(1.58)
Bootstrap	24.18	1.38	43.22	55.19	0.58	9.25	21.43
	(1.01)	(0.16)	(3.51)	(1.06)	(0.02)	(0.18)	(2.30)

NASDAQ

	γ	δ	σ	k	a	b	π
Baseline	25.68	1.20	38.60	63.12	0.58	9.17	23.79
	(0.78)	(0.10)	(3.37)	(1.13)	(0.01)	(0.17)	(2.39)
Bootstrap	24.18	1.38	43.22	55.19	0.58	9.25	21.43
	(1.01)	(0.16)	(3.51)	(1.06)	(0.02)	(0.18)	(2.30)

RUSSELL 2000

	γ	δ	σ	k	a	b	π
Baseline	24.44	1.07	36.70	64.70	0.64	8.34	26.02
	(0.60)	(0.09)	(3.95)	(2.13)	(0.01)	(0.15)	(2.28)
Bootstrap	24.18	1.38	43.22	55.19	0.58	9.25	21.43
	(1.01)	(0.16)	(3.51)	(1.06)	(0.02)	(0.18)	(2.30)

Table 9: Estimated parameters by maximum likelihood, for the security industry
Values for γ, σ are given as annualized percentages, and values for k, π are given as percentages. Standard errors are in parenthesis.

S&P500

	E[ln R]	E[R]	α	β
Baseline	31.82	43.40	36.10	1.56
	(43.32)	(48.59)	-	-
Bootstrap	30.20	42.74	35.67	1.51
	(44.60)	(50.19)	(3.63)	(0.34)

NASDAQ

	E[ln R]	E[R]	α	β
Baseline	30.99	40.85	34.66	1.31
	(39.77)	(44.27)	-	-
Bootstrap	30.20	42.74	35.67	1.51
	(44.60)	(50.19)	(3.63)	(0.34)

RUSSELL 2000

	E[ln R]	E[R]	α	β
Baseline	29.23	38.02	32.48	1.16
	(37.67)	(41.62)	-	-
Bootstrap	30.20	42.74	35.67	1.51
	(44.60)	(50.19)	(3.63)	(0.34)

Table 10: Implied estimates for $\mathbb{E}[\ln R]$, $\mathbb{E}[R]$, α and β for the security industry
Implied estimates for the expected value and standard deviation for returns and log-returns, as well as α and β. Values given as annualized percentages (except for β). Standard deviations are in parenthesis.

5. Conclusion

5.1. Limitations

The post-money valuation model is heavily limited as it truly only relies on the amount of money raised. Although the latter variable is definitely a key component of a start-up valuation, more meaningful parameters should be taken into account, such as accounting data, market projections, or technology prospects, which are not widely available. Firm valuation remains a discretionary task achieved by specialized analysts, and a systematic method can only give a rough estimate. The model does not try to estimate the true firm value, but rather the value from a VC perspective. Since the training data comes from VCs themselves, the model can only be as good as VCs are, with the same biases. Although estimates for log-valuations are fairly good, getting a correct estimate of the actual value of the firm is much more difficult due to the wide range of values that projects can take, with more than four order of magnitude differences.

The core methodology used in this research suffers from several limitations, mostly identical to those of the original paper. First, the iterative design of the simulation makes computations slow. Given the size of the dataset (close to 120,000 observations), the optimization process takes a few minutes to reach a satisfying convergence. This becomes a real issue to conduct a thorough bootstrap estimation, as the simulation itself cannot be parallelized. The solution would be to run several instances of the program on multiple cores and average out results. The time-grid is also set to three month, which is a good trade-off between speed and accuracy, but a one month time-step would still be interesting to investigate on a smaller sample. This latter implementation would probably need even more data to be able to accurately fit the monthly distributions of exits and closings.

The selection function remains in its simplest form, namely a function of firm value only, and is the same for any fate. It may be useful to distinguish different selection function for different fates, as the probability of going public surely does not depend on the same parameters as the probability of going out of business. This would further slow down the simulation process, as additional parameters would be needed to compute the probabilities.

Although the log-returns are fairly well described by a log-normal model, the equations do not capture any cross-correlation between project returns, which certainly exist. Parameters are also independent of the firm value, and it is likely that the slope δ is linked to the project maturity. The closer it is to exit, the more sensitive to the market it should be.

5.2. Further research

Further research can be conducted using an even more complete dataset. Although Crunchbase aggregates several data sources, it is far from complete. Preliminary investigations showed that Crunchbase misses entries found in other databases such as S&P Capital IQ.[6] The precise amount of missing data is yet to be determined. This task requires a substantial amount of time and careful processing to match databases entries, and merge them properly. In addition to that, none of these databases are free to access. Another important bottleneck is the exit value of firms. In case of an IPO, the value is often observable but for acquisitions a lot of data points are missing, leading to the impossibility to compute an actual return from investment to exit, even with the PMVs available.

This work could be further extended using more complex log-returns models such as the Johnson's S_U-distribution. This distribution is a better fit to the log-return process and has been used successfully to model asset returns for portfolio management and in option pricing. Empirically, we find that this is also the best fitting distribution for our dataset, among the several dozen distributions tested. The process being Gaussian, it is still fairly simple to implement, but also more costly in terms of computational time since it is a four-parameter distribution. Last, and as discussed previously, implementing a value-dependent parametrization could lead to even better estimates, as well as more complex selection functions.

5.3. Conclusion

In this paper we implement the maximum likelihood methodology first proposed by John H. Cochrane in his 2005 paper "The risk and return of venture capital", using a state-of-the-art dataset, Crunchbase. To alleviate the numerous missing observations, we develop a machine learning procedure using gradient boosting models to infer post-money valuations of VC-backed companies. This data augmentation allows to vastly increase the number of exploitable observations and further improve the original methodology.

The results show a fairly good log-valuation regression with less than 4% median average error. The best performing family of models is the (histogram-based) gradient boosting models. The main feature explaining post-money valuations according to the model is by far the amount of money raised. Being able to infer with good precision these log valuations allows to compute log-returns and fully leverage the database.

The results show a very high positive α for all market sectors and especially the cybersecurity sector. This confirms the findings of previous works, that find similar values. Overall

[6] S&P Capital IQ Pro, data as of July 1, 2021

the private equity market is riskier than its public counter part, exhibiting values for β greater than 1, except for specific sectors such as the health industry, for which β is lower than 1. Venture capital risk arises mainly from the high idiosyncratic risk, as most project either fail or remain private and do not exit, yielding no return.

Although these results are not sufficient at all to dictate any investment decision, they allow to get a picture of the venture capital market for the past 10 years. It is part of a wider effort by the CYD Campus and the TMM team to assess potential strengths and opportunities in the cyber-security industry. With sufficient data, this work can be more specifically applied to smaller parts of the market such as the Swiss venture capital market. Combined with other tools such as the "TechRank" approach (**?**), it would contribute to guide and help investors to undertake a transparent decision when dealing with highly complex scenarios.

References

Alexy, O. T., Block, J. H., Sandner, P., Ter Wal, A. L. J., 2012. Social capital of venture capitalists and start-up funding. Small Business Economics 39, 835–851.

Ang, A., Chen, B., Goetzmann, W. N., Phalippou, L., 2018. Estimating private equity returns from limited partner cash flows. Journal of Finance 73, 1751–1783.

Axelson, U., Martinovic, M., 2015. European venture capital: Myths and facts. Available at `https://personal.lse.ac.uk/axelson/ulf_files/EuroVC_MythsFacts%20v17.pdf`

Besten *den*, M. L., 2021. Crunchbase research: Monitoring entrepreneurship research in the age of big data. Available at `http://dx.doi.org/10.2139/ssrn.3724395`

Chernenko, S., Lerner, J., Zeng, Y., 2021. Mutual funds as venture capitalists? Evidence from unicorns. Review of Financial Studies 34, 2362–2410.

Cochrane, J. H., 2005. The risk and return of venture capital. Journal of Financial Economics 75, 3–52.

Cumming, D., Haß, L. H., Schweizer, D., 2013. Private equity benchmarks and portfolio optimization. Journal of Banking and Finance 37, 3515–3528.

Dalle, J.-M., Besten *den*, M. L., Menon, C., 2017. Using Crunchbase for economic and managerial research. Available at `https://doi.org/10.1787/18151965`

Driessen, J., Lin, T.-C., Phalippou, L., 2012. A new method to estimate risk and return of nontraded assets from cash flows: The case of private equity funds. Journal of Financial and Quantitative Analysis 47, 511–535.

Ewens, M., 2009. A new model of venture capital risk and return. Available at `http://dx.doi.org/10.2139/ssrn.1356322`

Feurer, M., Eggensperger, K., Falkner, S., Lindauer, M., Hutter, F., 2021. Auto-Sklearn 2.0: Hands-free AutoML via meta-learning. Available at `https://arxiv.org/abs/2007.04074`

Florackis, C., Louca, C., Michaely, R., Weber, M., 2022. Cybersecurity risk. Review of Financial Studies, forthcoming.

Franzoni, F., Nowak, E., Phalippou, L., 2012. Private equity performance and liquidity risk. Journal of Finance 67, 2341–2373.

Gornall, W., Strebulaev, I. A., 2020. Squaring venture capital valuations with reality. Journal of Financial Economics 135, 120–143.

Harris, R. S., Jenkinson, T., Kaplan, S. N., 2016. How do private equity investments perform compared to public equity? Journal of Investment Management 14 (3), 14–37.

Hervé, F., Schwienbacher, A., 2018. Round-number bias in investment: Evidence from equity crowdfunding. Finance 39 (1), 71–105.

Hwang, M., Quigley, J. M., Woodward, S. E., 2005. An index for venture capital, 1987–2003. Contributions to Economic Analysis and Policy 4, 1–43.

Korteweg, A., Nagel, S., 2016. Risk-adjusting the returns to venture capital. Journal of Finance 71, 1437–1470.

Korteweg, A., Sorensen, M., 2010. Risk and return characteristics of venture capital-backed entrepreneurial companies. Review of Financial Studies 23, 3738–3772.

Kwon, S., Lowry, M., Yiming, Q., 2020. Mutual fund investments in private firms. Journal of Financial Economics 136, 407–443.

Moskowitz, T. J., Vissing-Jørgensen, A., 2002. The returns to entrepreneurial investment: A private equity premium puzzle? American Economic Review 92, 745–778.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peng, L., 2001. Building a venture capital index. Available at `http://dx.doi.org/10.2139/ssrn.281804`

Phalippou, L., 2009. Beware of venturing into private equity. Journal of Economic Perspectives 23 (3), 147–166.

Schmidt, D. M., 2006. Private equity versus stocks. Journal of Alternative Investments 9 (1), 28–47.