Risk and Return Characteristics of Venture Capital-Backed Entrepreneurial Companies

Arthur Korteweg

Morten Sorensen[†]

August 2009

Abstract: Valuations of entrepreneurial companies are only observed occasionally, albeit more frequently for well-performing companies. Consequently, estimators of risk and return must correct for sample selection to obtain consistent estimates. We develop a general model of dynamic sample selection and estimate it using data from venture capital investments in entrepreneurial companies. Our selection correction leads to markedly lower intercepts and higher estimates of risks compared to previous studies.

[†] Stanford University, GSB (korteweg@stanford.edu) and Columbia Business School and NBER (ms3814@columbia.edu). We are particularly grateful for our discussions with John Cochrane that helped shape our understanding of the underlying problems, and we thank John Heaton, Josh Lerner, John Quigley, Caroline Sasseville, Annette Vissing-Jørgensen, and seminar participants at the University of Amsterdam, Columbia University, University of Chicago, University of Illinois–Urbana Champaign, Copenhagen University, London Business School, the Stanford-Berkeley joint seminar, Tilburg University, the University of Iowa and the 2009 American Finance Association Meetings in San Francisco for helpful discussions. Susan Woodward of Sand Hill Econometrics provided generous support with access to her data, and the Center for Research in Security Prices (CRSP) and the Kauffman Foundation provided financial support.

There are many assets that only trade infrequently, such as privately held companies, real estate, corporate and municipal bonds, small-cap stocks, many structured products, and securities trading OTC. Since their valuations are only known when they trade, valuation and return data for these assets are necessarily sporadic. It is well known that when assets trade nonsynchronously and are "kept on the books" at previous trading prices, the stale price problem biases estimates of risk and return (Scholes and Williams (1977) and Dimson (1979)).[1] Moreover, when the timing of the observed returns is endogenous, a sample selection problem arises. Here we address the latter problem, which we term the *dynamic selection problem*.

The dynamic selection problem is widespread. For hedge funds, a number of studies identify the selection issues that arise from the nature of hedge fund data (e.g., Baquero, ter Horst, and Verbeek (2005) and Jagannathan, Malakhov, and Novikov (2009)). Hedge fund data are based on voluntary performance reports, and hedge funds with worse performance may, naturally, be more reluctant to report their returns and less likely to survive. The resulting self-selection and survivorship problems are manifestations of the dynamic selection problem. Another example is real estate, where prices are only observed for properties that trade, and these properties may not be representative of the overall housing stock. The dynamic selection problem would arise if, say, sellers with higher reservation prices are less likely to sell (Hwang and Quigley (2003) and Goetzmann and Peng (2006)) or if properties that have depreciated more are more likely to trade, e.g. due to foreclosures. Similarly, for venture capital and private

[1] Chapter 3 in Campbell, Lo, and MacKinlay (1997) presents a textbook discussion of the econometric issues arising from nonsynchronous trading and references to the extensive literature on this topic.

equity investments, the valuations of portfolio companies are only known when they receive funding or have exit events, i.e. go public or are acquired. These events are more frequent for well-performing companies, and these companies are more frequently observed in the data, giving rise to the dynamic selection problem (Cochrane (2005) and Hwang, Quigley, and Woodward (2005)).

Our study focuses on venture capital investments in entrepreneurial companies. Controlling for selection, we find significant decreases in the intercept and alpha of the market model and substantial increases in risk measures. In our baseline specifications, the estimated alpha decreases by about 40% and the market beta increases by about 20%, compared to GLS estimates that ignore the selection problem. In a three-factor specification (Fama and French (1995)), the loading on the size factor turns positive, and the loading on the book-to-market factor (HML) increases by about 70%, resulting in factor loadings on the size factor (SMB) of about 1.1 and on the book-to-market factor of about -1.6. Perhaps not surprisingly, the risk profile of privately held entrepreneurial companies resembles that of small, high-risk, and very high-growth public companies.

Our econometric approach combines a selection process with a standard dynamic asset-pricing model. We embed a Type-2 Tobit model (Heckman (1979) and Amemiya (1985)) into a dynamic filtering and smoothing problem (Kalman (1960) and Anderson and Moore (1979)). The model explicitly specifies the entire valuation path between the observed valuations as well as the probability of observing a valuation at each intermediate point in time. Our selection process is a function of the contemporaneous valuation, various measures of market conditions, and the time elapsed since the previous

refinancing (non-linearly). The time since the previous refinancing round is important for the identification of the model. It is well known (Heckman (1990) and Andrews and Schafgans (1998)) that semi-parametric identification of sample selection models requires independent variation in the selection equation. Without such variation the estimated parameters may be sensitive to functional and distributional assumptions. As argued below, the time since the previous refinancing is a reasonable source of such variation, and we confirm in Appendix B that our results are robust to relaxing the distributional assumptions.

For estimation, we adopt a Bayesian approach. This approach has several advantages over frequentist alternatives, such as maximum likelihood. First, the Bayesian approach is numerically more tractable than existing methods, allowing us to estimate more flexible specifications and more general error distributions. Second, the Bayesian approach delivers accurate finite-sample inference, even for non-linear functions of non-Gaussian parameters, such as the alpha in our model. Moreover, although we primarily report point estimates, the procedure generates the entire posterior distribution of all the parameters of interest as well as the latent variables in the model. Finally, our approach does not require restricting the valuations to a grid, and it preserves their full continuous distributions. To estimate our model, we use a Markov Chain Monte Carlo algorithm called Gibbs sampling (Gelfand and Smith (1990) and Robert and Casella (2004)). The algorithm produces the posterior distribution by iteratively simulating from three simpler distributions: a Bayesian regression, a draw of truncated random variables, and a path from a Kalman Filter. Each of these simpler distributions is well understood and

tractable, and combined they form an estimation procedure that is surprisingly manageable given the numerical complexity of the model.

A. The Dynamic Selection Problem

To fix ideas and notation, the generic *dynamic selection model* consists of an outcome equation

$$v(t) = v(t-1) + X'(t)\theta + \varepsilon(t), \tag{1}$$

where $v(t)$ is the (log-)valuation at time *t*, and θ contains parameters of interest. The valuation is only observed when

$$w(t) \geq 0, \tag{2}$$

where $w(t)$ is a latent selection variable given by the selection equation

$$w(t) = Z'(t)\gamma_0 + v(t)\gamma_v + \eta(t). \tag{3}$$

Assuming $\varepsilon(t) \perp \eta(t)$ and $E[\varepsilon(t)] = 0$, the sample selection problems arises when $\gamma_v \neq 0$, because $E[\varepsilon(t) \mid data] \neq 0$, conditioning on all observed data.

In the standard cross-sectional case, *without* $v(t-1)$ in the outcome equation, a common two-step approach is to first calculate $E[\varepsilon(t) \mid data]$ and then include it as an additional variable (a *control function*) in the outcome equation. With Normal distributed errors and the standard normalization $\sigma_\eta = 1$ this conditional mean admits a closed form expression (see Heckman (1979)):

$$E[\varepsilon(t) \mid data] = E[\varepsilon(t) \mid w(t) \geq 0, Z(t), X(t)] = \frac{\gamma_v \sigma_\varepsilon^2}{\sqrt{\gamma_v^2 \sigma_\varepsilon^2 + 1}} \frac{\phi(C(t))}{\Phi(C(t))}, \tag{4}$$

where $C(t) = \left(Z'(t)\gamma_0 + X'(t)\theta\gamma_v \right) \Big/ \sqrt{\gamma_v^2 \sigma_\varepsilon^2 + 1}$.

In contrast our dynamic model, *with* $v(t-1)$ in the outcome equation, is more complex. Consider a company that trades twice, at times t^0 and t^1, and hence only $v(t^0)$ and $v(t^1)$ are observed. Iterating on equation (1) yields:

$$E\left[v(t^1) \middle| data \right] = v(t^0) + \left[\sum_{\tau=t^0+1}^{t^1} X'(\tau) \right] \theta + E\left[\sum_{\tau=t^0+1}^{t^1} \varepsilon(\tau) \middle| data \right]. \tag{5}$$

The first term in brackets is a linear function of observed variables during the interim period. The last term is the error term, but its conditional mean is now

$$E\left[\sum_{\tau=t^0+1}^{t^1} \varepsilon(\tau) \mid data \right] = \left[\sum_{\tau=t^0+1}^{t^1} \varepsilon(\tau) \middle| \begin{array}{l} w(t^0) \geq 0, w(t^0+1) < 0, \ldots, \\ w(t^1-1) < 0, w(t^1) \geq 0, \\ Z(t^0), \ldots, Z(t^1), X(t^0), \ldots, X(t^1) \end{array} \right] \tag{6}$$

Note that the conditional mean is a function of the trading history and the observables over the entire period between the observed valuations. Unlike the standard selection model, observations with unobserved valuations are still informative about the valuation process, and the conditional means for the observed valuations also depends on the periods where the valuations were unobserved. Further, the conditional mean depends on the trading history, and $E\left[\sum_{\tau=t^0+1}^{t^1} \varepsilon(\tau) \middle| data \right]$ will differ for a company that is observed

to trade twice, at times $t^0 < t^1$, and another company that is observed three times at times $t^0 < t' < t^1$. Accounting for these dependencies is difficult, however, because it requires integrating over all possible paths of the unobserved outcome and selection processes, and reasonable specifications typically lead to intractable models.

In the next section we present the related literature. Section II describes our econometric model and estimation algorithm, and section III describes our data. Section IV discusses our empirical results. In section V we discuss the interpretation of the intercepts in our factor models, and section VI concludes. The two appendices contain additional technical details: Appendix A describes the estimation algorithm in detail, and appendix B describes the robustness and convergence properties of this algorithm.

I. Related Literature

The dynamic selection problem is encountered in a number of studies. Real estate repeat-sales indices, such as the S&P/Case-Shiller index, suffer from this problem, as discussed in Gatzlaff and Haurin (1997), Fisher, Gatzlaff, Geltner, and Haurin (2003), Hwang and Quigley (2003), and Goetzmann and Peng (2006). To address the problem, these studies estimate variations of the cross-sectional Heckman specification in equation (4). While this specification is tractable, it is unattractive for two reasons: First, it ignores the conditioning on intermediate periods from equation (6), so it may be misspecified, raising concerns about the studies' ability to accurately capture the selection dynamics and produce consistent estimates. Second, it only permits selection based on a property's contemporaneously observable characteristics, such as its age and size, but not its price

appreciation or depreciation over a period of time. Hence, these studies may control for bias arising if sellers with greater reservation prices are less likely to trade or trade at higher prices, which is how they are motivated. But they cannot capture the bias that would arise if properties that have depreciated more are more likely to trade, e.g. due to foreclosures, a potentially important concern.[2] Inspired by this literature, Hwang, Quigley, and Woodward (2005) use this approach to construct a venture capital index, sharing the same limitations as the real estate studies.

In the hedge fund literature a number of studies have investigated the implications of the various selection issues inherent in hedge fund data. Hedge fund databases, such as TASS and HFR, are constructed from voluntary performance reports, but funds with worse performance are less likely to report and subsequently survive, creating self-selection and survivorship biases. Baquero, ter Horst, and Verbeek (2005) and ter Horst and Verbeek (2007) present propensity-weighting procedures to correct for self-selection and survivorship biases. More recently, Jagannathan, Malakhov, and Novikov (2009) study persistence of returns and control for selection using GMM estimation. These two papers assume that selection depends only on observable characteristics, such as investment style, age, and past observed returns, but *not* contemporaneous, potentially unobserved, performance. Hence, these specifications may capture the survivorship bias that would arise when poor observed performance over a number of prior periods leads to a fund's eventual demise. However, to capture self-selection arising from fund managers'

[2] The New York Times article "Where Housing Crashed Early, Glimmers of Recovery Emerge" (May 5, 2009) reported that two-thirds of all sales in Sacramento, CA, during March 2009 were bank repossessions.

unwillingness to volunteer information about poor current performance, it seems necessary to include contemporaneous performance in the selection equation as well.

Hence, a limitation of these studies is that the selection equation does not depend on the contemporaneous, potentially unobserved, valuations. This problem is addressed by Cochrane (2005), who lets the selection equation depend on this contemporaneous valuation, but nothing else. In his study, which is closely related to ours, he uses a ML procedure to estimate the risk and return of venture capital investments. This approach suffers from numerical difficulties associated with estimating the conditional mean of the error terms.[3] His model estimates seven parameters, leading to somewhat parsimonious specifications of both the selection and outcome equations.

We extend Cochrane's approach in a number of ways, resulting in markedly different estimates. First, we include the time since previous refinancing and various market conditions, providing more reasonable and arguably better identified specifications, and we find strong and significant coefficients on these variables. We estimate more flexible specifications of the outcome equation with more factors and separate coefficients for companies at different stages and over different periods. Second, our approach does not require restricting the valuations to a grid, but maintains their full continuous distributions. Third, our algorithm produces accurate finite sample inference. Cochrane (2005) reports bootstrapped standard errors that are consistently an order of magnitude greater than the asymptotic ones, suggesting that this is not a trivial concern.

[3] For example, in Cochrane's footnote 1 he mentions that "a smooth [selection equation] would be prettier, but this specification requires only one parameter, and the computational cost of extra parameters is high."

Finally, we estimate our model with an updated dataset, and at the monthly, not quarterly, frequency. Combined, these factors lead to substantially different results.

The main difference is our estimates of systematic risks. Cochrane's round-to-round estimates, which are most comparable to ours, show betas that are consistently below 1.0, with an average beta of just 0.6. In contrast, we find betas that are consistently above 2.2, with an average beta of 2.8. This difference is substantial. To contrast with previous findings that do not correct for selection and hence may underestimate the betas, Reyes (1990) finds betas ranging from 1.0 to 3.8 (using data from 175 mature venture capital funds), and Gompers and Lerner (1997) report betas from 1.08 to 1.4 (using a sample of 96 venture capital investments). Peng (2001), using a propensity weighting method, reports betas ranging from 1.3 to 2.4 on the S&P 500 and from 0.8 to 4.7 on NASDAQ. Further we calculate the loadings of entrepreneurial firms on the size and book-to-market factors (Fama and French (1995)). We find that returns to entrepreneurial companies behave similarly to the returns to small high-growth companies, their publicly traded counterparts. In addition to the different betas, we also find higher idiosyncratic risk and lower intercepts. Our idiosyncratic volatility estimate is around 130% per year, compared to Cochrane's estimate of about 84%.

Another related literature estimates the risk and return of private equity and venture capital investments using the cash flows distributed by the funds to their limited partners (Gompers and Lerner (1997), Jones and Rhodes-Kropf (2003), Ljungqvist and Richardson (2003), Kaplan and Schoar (2005), Phalippou and Gottschalg (2005), and Driessen, Lin, and Phalippou (2007)). One limitation of this approach is that the return to

a fund is earned across a portfolio of companies, typically over a ten- to thirteen-year period, making it difficult to use fund level returns to identify differences across shorter time periods, across industries, and across companies with different characteristics, such as their stage of development. Estimation using valuations of individual companies may provide a more nuanced view of these differences. Moreover, using individual valuations leads to substantially more independent observations and consequently greater statistical power.

II. Econometric Model

To motivate the specifications and help interpret and compare the results to OLS and GLS estimates, we first derive the discrete-time valuation process from a continuous-time specification

A. The Valuation Process

Let the economy contain a risk-free bond with price $B(t)$, paying the continuously compounded rate r

$$\frac{dB(t)}{B(t)} = rdt\,. \tag{7}$$

The value of the market portfolio, $M(t)$, follows a geometric Brownian motion

$$\frac{dM(t)}{M(t)} = \mu_m dt + \sigma_m dW_m(t)\,, \tag{8}$$

where μ_m is the drift, and $W_m(t)$ is a Wiener process. The valuation of a given company is $V(t)$, and it develops according to the one-factor market model

$$\frac{dV(t)}{V(t)} - rdt = \alpha dt + \beta\left(\frac{dM(t)}{M(t)} - rdt\right) + \sigma dW(t). \tag{9}$$

The drift of the valuation process in excess of *r* is α, and $dW(t)$ is independent of $dW_m(t)$ per definition of beta. Denote the continuously compounded returns $r_v(t,t') = \ln[V(t')/V(t)]$ and $r_m(t,t') = \ln[M(t')/M(t)]$, and define $\delta = \alpha - \frac{1}{2}\sigma^2 + \frac{1}{2}\beta(1-\beta)\sigma_m^2$. Using Itô's lemma, we derive the discrete-time return equation

$$r_v(t,t') - (t'-t)r = (t'-t)\delta + \beta\left(r_m(t,t') - (t'-t)r\right) + \varepsilon(t,t'), \tag{10}$$

where $\varepsilon(t,t')$ is distributed $N(0,(t'-t)\sigma^2)$. Defining $v(t) = \ln[V(t)]$, and starting from $t = t'-1$, we arrive at the one-period transition equation for the outcome equation

$$v(t) = v(t-1) + r + \delta + \beta\left(r_m(t) - r\right) + \varepsilon(t), \tag{11}$$

with $\varepsilon \sim N(0,\sigma^2)$ and $r_m(t) = \ln[M(t)/M(t-1)]$.[4] This is equation (1) with $X'(t) = [1 \quad r_m(t) - r]$ and $\theta = [r + \delta \quad \beta]'$.

[4] For multi-factor models, $\delta = \alpha - \frac{1}{2}\sigma^2 + \frac{1}{2}\beta' diag(\Sigma) - \frac{1}{2}\beta'\Sigma\beta$ where Σ is the covariance matrix of the factor returns.

B. The Selection Process

Valuations are only observed when a company has a refinancing or an exit event, and the endogeneity of these events is captured by the selection process. Following equations (2) and (3), let $v(t)$ be observed only when

$$w(t) \geq 0 , \tag{12}$$

where $w(t)$ is a latent selection variable specified as

$$w(t) = Z'(t)\gamma_0 + v(t)\gamma_v + \eta(t) . \tag{13}$$

The vector $Z(t)$ contains characteristics that affect refinancing and exit events, including a constant term, the time since the previous financing round (linearly and squared), and variables capturing general market conditions. The second term in equation (13) is the log-valuation. By including the log-valuations at the previous financing round in $Z(t)$ with a coefficient of $-\gamma_v$, we can interpret γ_v as the coefficient on the return earned since this previous round. Since valuations are observed more frequently for more successful companies, we expect γ_v to be positive. As usual for selection models, the scale of the selection equation is unidentified and is normalized by fixing the variance of the error term to equal one. Hence, we assume $\eta(t)$ is distributed *i.i.d.* $N(0,1)$.

To summarize, the model contains two equations: the valuation equation (11) and the selection equation (13). Only when $w(t) \geq 0$ is $v(t)$ observed, and $w(t)$ is never

observed. The error terms are distributed *i.i.d.* $\varepsilon(t) \sim N(0,\sigma^2)$ and $\eta(t) \sim N(0,1)$. The parameters of interest are δ, β, σ^2, and $\gamma = (\gamma_0, \gamma_v)$.

C. Overview of Estimation Procedure

Appendix A presents the estimator in detail. We use a Bayesian Gibbs sampling procedure (see Geman and Geman (1984), Tanner and Wong (1987), Gelfand and Smith (1990) and Johannes and Polson (2006)), which allows us to divide our model into three blocks: The first one contains the valuation variables, the second contains the selection variables, and the last block contains the parameters of interest. The Gibbs sampler simulates the joint (augmented) posterior distribution of the model by iteratively sampling the variables in each block conditional on the previous realizations of the variables in the other blocks. The second and third blocks are simple: For the selection variables in the second block, we sample from truncated Normal distributions, defined by equations (12) and (13). This step is similar to Bayesian estimation of Probit models (Albert and Chib (1993)). For the parameters in the valuation and selection equations in the third block, we use two standard Bayesian linear regressions, following equations (11) and (13).

Drawing the valuation variables in the first block is the most complex part of the algorithm. The procedure must trace out the entire path of the unobserved valuations, conditioning on the parameters, selection variables, market returns, and on the fact that during this intermediate period no valuations were observed, which shifts down the unobserved valuations' conditional distributions. We use the Forward Filtering

Backwards Sampling (FFBS) procedure by Carter and Kohn (1994) and Fruhwirth-Schnatter (1994), which provides an efficient way to sample a path of latent variables conditional on all available information.

To understand the application of the FFBS procedure, note that conditional on the parameters and selection variables, the model is a linear state space, and the path of the latent valuations can be recovered using a Kalman filter. From this perspective, $v(t)$ are unobserved state variables, and the valuation equation (11) is the transition rule with $r + \delta + \beta(r_m(t) - r)$ as an "observed" control acting on the state. The state space has one or two observation equations depending on whether the valuation is observed or not: Conditionally, $w(t)$ can be viewed as noisy "observations" of $v(t)$, and the first observation equation is the selection equation (13). Second, when a valuation is observed, it provides a direct observation of the underlying state and $\ln[V_{OBS}(t)] = v(t)$, where $V_{OBS}(t)$ is the observed valuation as defined in the next section. We assume that valuations are observed without error, although it is possible to incorporate observation error without losing the linear filtering properties.

We use diffuse priors and several different starting values for the parameters, as detailed in Appendix A. Our Gibbs sampler uses 1,000 iterations for the initial burn-in followed by 5,000 iterations to simulate the posterior distribution. During the burn-in, the simulations converge quickly. We verify the convergence and robustness of the algorithm in Appendix B, including relaxing the assumption of Gaussian error terms.

III. Data Description

Monthly market returns and returns to Fama-French portfolios are taken from Kenneth French's website. These are constructed from the NYSE, AMEX, and NASDAQ firms in CRSP. Monthly Treasury-Bill rates are from Ibbotson Associates and are also available on this website.

A. Venture Capital Data

Venture capital investment data were provided by Sand Hill Econometrics (SHE). SHE combines and extends two commercially available databases: VentureXpert (formerly Venture Economics) and VentureSource (formerly Venture One). These two databases are used extensively in the VC literature, and the combined data contain the majority of US VC investments from 1987 to 2005. Gompers and Lerner (1999) and Kaplan, Sensoy, and Strömberg (2002) investigate the completeness of VentureXpert and find that missing investments are predominantly smaller and more idiosyncratic ones. In addition, SHE has spent a substantial amount of time and effort to ensure the accuracy of the data. This includes removing duplicate investment rounds, adding missing rounds, and consolidating rounds, ensuring that each round corresponds to a single investment by one or more VCs. Cochrane (2005) uses an earlier version of these data, and the previously reported data problems have been resolved.[5]

[5] Cochrane reports that, in his version of the dataset, liquidation dates were unreliable and apparently clustered on two specific days prior to 1997 (not accounting for this clustering led to negative estimates of betas). Moreover, he explicitly models measurement error and filter the data to account for outliers. In contrast, we only had to eliminate a single round in which the return was < -100%. In Appendix B we relax

B. Calculating Returns

VCs distinguish between pre- and post-money valuations. To illustrate, when a VC invests I in a company with a total valuation of V_{POST} (the post-money valuation), V_{PRE} (the pre-money valuation) is defined by $V_{POST} = V_{PRE} + I$. Hence, we calculate the return earned by an investor over two subsequent rounds from time t to t' as

$$R_v(t,t') = V_{PRE}(t')/V_{POST}(t). \tag{14}$$

We use these returns to construct a new valuation variable, which strips out the effects of ownership dilution by future investors. Starting from $V(0) = 1$, the dilution-adjusted valuations are calculated iteratively as

$$V_{OBS}(t') = V_{OBS}(t) \times R_v(t,t'). \tag{15}$$

These valuations are used as the observed valuations in the estimation procedure.

This calculation requires valuations that are observed for consecutive rounds. When a valuation is missing for an intermediate round, it is not possible to adjust for dilution, and the dilution-adjusted valuation is restarted after the break in observed valuations. For firms that are liquidated at an unknown amount, we set the liquidation value equal to 10% of the original investment.[6]

---

the Normality assumption to allow for fat tails and skewed distributions, and we find no evidence of outliers in our data.

[6] Our results are not sensitive to this assumption. In our base specification, we estimate an intercept of -0.0563 and a beta of 2.7510. With a liquidation rate of 25%, the intercept changes to -0.0566 and the beta becomes 2.7900. The coefficients in the selection equation are similarly unaffected.

C. Descriptive Statistics

The full dataset contains 61,356 investment rounds for 18,237 companies. However, we only have valuation data for a fraction of these companies. Moreover, private financing rounds are less public events than IPOs and acquisitions, hence companies that end up going public or being acquired are overrepresented in the subsample for which valuations are available. Consequently, we use a random procedure to scale the number of companies in our sample to match the exit rates in the full dataset and ensure that these data are representative.[7] The number of companies and their exits, in the full dataset and in our sample used for estimation, are listed in Table I.

**** TABLE I: DESCRIPTIVE STATISTICS ****

Our final sample contains a total of 5,501 financing rounds for 1,934 companies. Of these, 199 (10.3%) companies go public, another 451 (23.3%) are acquired, and we have information that 445 (23.0%) have been liquidated. We have no information about the fate of the remaining 839 (43.4%) companies. Some of these may be alive and well, some may be "living zombies," but the majority has likely been liquidated at this point. The empirical model incorporates the uncertainty about these unobserved outcomes by simulating valuations for 60 months past the last observed round.[8]

[7] Here and below, we assume that the subsample of rounds for which we have valuations is a random subsample (conditional on the observed exit). If not, this would potentially create an additional sample selection problem. Using the full sample of observed valuations raises the intercept of the outcome equation by about 1%/month, but has negligible effect on the other parameter estimates.

[8] Our results are robust to this assumption. Extending the period to 120 months, the base estimates of -0.0563 and 2.7510 decrease to -0.0625 and 2.6806 for the intercept and beta, respectively. The coefficients in the selection equation are even less affected.

An entrepreneurial firm receives 4.4 financing rounds on average (the median is 4 rounds), with some firms receiving as many as 9 rounds. On average, 13 months pass between rounds (the median is 10 months). While 5% of follow-on investments occur after as few as 2 months, another 5% take 34 months or more. The average arithmetic return between observed rounds is 95% (median 21%) with a standard deviation of 319%.

IV. Risk Factors for Entrepreneurial Companies

Table II presents four different specifications of the selection equation. The valuations appear highly exposed to the market factor with a beta coefficient (RMRF) around 2.8. The (monthly) intercept is about -5.7%, and the (monthly) standard deviation of the idiosyncratic returns (Volatility) is 41%. These coefficients are stable across specifications. For comparison, Davis, Fama, and French (2000) consider companies trading on NYSE, AMEX, and NASDAQ, and for small growth companies – most similar to our entrepreneurial ones – they estimate betas from 1.01 to 1.06, depending on the time period, considerably less than betas for the companies in our sample.

**** TABLE II: ONE-FACTOR MODEL ****

In the first specification, the selection equation includes only the (log-)return and the time since the previous financing round, linearly and squared. The coefficient on the return is positive and highly significant across all specifications. Companies with higher returns are more likely to have refinancing or exit events and hence appear in the data, suggesting that the sample selection problem may be substantial.

The coefficients on time and time squared are around 0.4 and -0.04. This captures the distribution of the frequency of refinancing rounds. Keeping the valuation constant, the probability of observing a refinancing or exiting event each month (a hazard rate) increases from the time of the previous round and reaches a maximum after roughly 5 years $(= 0.4/(2 \times 0.04))$ after which the likelihood decreases. The negative square term captures the rapid deterioration of the likelihood of refinancing and the corresponding higher returns required to achieve them as more time passes. This captures the fact that companies that have not received financing for a while become increasingly unlikely to ever receive refinancing again.

Semi-parametric identification of selection models requires a variable that enters the selection equation but is independent of the error term in the valuation equation (Heckman (1990) and Andrews and Schafgans (1998)). The time since the previous refinancing round (*Time*) seems a reasonable source of such variation. Given the standard assumption that valuations incorporate all contemporaneous information, the error terms in the valuation equation are independent over time. In particular, next period's error term is independent of the current *Time*, and hence it is independent of next period's *Time* as well. Moreover, it is reasonable to believe that *Time* is directly related to the probability of observing a financing round, since the probability of getting refinanced is low immediately after receiving a previous round, then it increases, and finally it declines after too much time has elapsed. The empirical results confirm this pattern. To illustrate, say a VC provides sufficient funding for companies to sustain themselves for 13 months exactly (our sample average). At this point, the company will only be refinanced if it is

performing well. Hence, an indicator variable that equals one after 13 months will be positively related to the probability of observing a refinancing. If the company is not observed to be refinanced at this point, it is struggling, not meeting the VC's threshold for reinvesting, and it will disappear. In this case, *Time* will be valid source of exogenous variation for semi-parametric identification of the model.

In the second specification of the selection process in Table II, the market return (RMRF) enters the selection equation with a negative coefficient. This may seem puzzling, but to derive the full effect of the market on the probability of observing a valuation, the indirect effect of the market on the valuation should also be considered. To illustrate, using the estimates in specification (2) in Table II, let RMRF increase by one. On average, this translates into an increase in the valuation of 2.79, and the combined effect on the selection equation is $2.79 \times 0.34 - 0.71 = 0.25$, which is positive, consistent with the empirical fact that more valuations are observed when the market is higher.

In addition to the return and the time since the previous financing round, there may be a cyclical component to VC investments – "hot" and "cold" markets – and the variables *Acquisitions*, *IPOs*, and *Rounds* control for this market cycle: *Acquisitions* contains the number of VC-backed acquisitions during the same month as the investment, *IPOs* contains the number of VC-backed IPOs during this month, and *Rounds* contains the number of investments rounds during this month. In the selection equation these are strongly significant, but they have little effect on the estimates in the valuation equation. It is surprising that *IPOs* enters with a negative sign, but this variable is correlated with the Acquisitions and Rounds variables.

Overall, the estimates of the valuation equation are robust across specifications, and the more parsimonious specification appears to capture the selection well. The richer specifications suggest that VC investments have a cyclical component that is not captured by the traditional risk factors, and we explore the role of this component further below.

A. Magnitude of Selection Bias

To assess the magnitude of the selection bias, we compare our estimates to OLS, GLS, and MCMC estimates that do not correct for selection bias. Table III presents estimates of these models, not correcting for selection. For the standard OLS and GLS estimators, we calculate the log excess returns and regress them on the corresponding log excess market returns. For the OLS estimator, we estimate the following specification, motivated by equation (10), pooled across firms:

$$r_v(t,t') - (t'-t)r = (t'-t)\delta_{OLS} + \beta_{OLS}\left(r_m(t,t') - (t'-t)r\right) + \varepsilon_{OLS}. \tag{16}$$

The coefficient δ_{OLS} corresponds to the one-period intercept in equation (11), and it is called *Intercept* here as well although, strictly speaking, equation (16) has no intercept. When the observed valuations are more distant in time they have more volatile errors, introducing heteroscedasticity. Ignoring selection, equation (10) implies that

$$\varepsilon_{OLS} \sim N\left(0,(t'-t)\sigma^2\right), \tag{17}$$

and the GLS estimator normalizes the variance of the error term by dividing by the square root of the time between observed valuations:

$$\frac{r_v(t,t')-(t'-t)r}{\sqrt{t'-t}} = \left(\sqrt{t'-t}\right)\delta_{GLS} + \beta_{GLS}\left(\frac{r_m(t,t')-(t'-t)r}{\sqrt{t'-t}}\right) + \varepsilon_{GLS}\,. \tag{18}$$

Again, δ_{GLS} corresponds to δ in equation (11) and is called *Intercept* here as well.

**** TABLE III: OLS, GLS, AND MCMC ****

Comparing the OLS and GLS estimates, we note that the OLS estimators have lower intercepts, corresponding to lower monthly drifts. The OLS estimators place relatively more weight on observations that are further apart, and the lower intercept indicates that these observations have lower average monthly returns than rounds that are closer together. This is not consistent with the observed valuations being generated by a standard Geometric Brownian motion, which has the same average monthly return regardless of the duration between rounds. However, as illustrated in Figure 1, it *is* consistent with the observations being generated by a selection process. Figure 1 illustrates a Geometric Brownian Motion with drift. The drift is indicated by the sloped solid line, and the process is observed when it is above a given threshold, illustrated by the horizontal line. The solid points represent the observed data points, and the gray points are unobserved ones. Point A represents an average observation after $t = 1/2$. Conditional on being observed at this point, the observations must have a high realized drift to make it across the threshold, as illustrated by the steep dotted line reaching this point. The point B represents the average observation at $t = 2$. Conditional on being observed at this point, the process needs a somewhat lower drift, on average, as indicated

by the flatter dotted line reaching point B. The finding that the OLS intercept is lower than the GLS intercept is consistent with this picture.

Like the GLS estimators, the MCMC specifications in table III also ignore the selection (by setting $\gamma_v = 0$, see details in appendix A). Comparing these MCMC specifications to the specifications in Table II with selection corrections, we find that the intercept declines from -1.6% to -5.7% per month. The change in the Beta (RMRF) is smaller, increasing from 2.66 without selection correction to 2.75 with correction, and the estimated volatility increases from a monthly standard deviation of 36% to 41%. These changes are all consistent with selected data, as illustrated in Figure 2. In this figure, the data are generated by a standard CAPM relationship, but they are only observed when the excess return is positive. Consistent with our empirical findings, the observed, selected observations in this figure have a flatter slope, a higher intercept, and a lower idiosyncratic volatility than the underlying true process.[9]

B. Three-Factor Model

Table IV presents estimates of a Fama-French three-factor specification, which includes the size (SMB) and book-to-market (HML) factors in addition to the market factor (RMRF). Again, we find substantial loadings on the market factor from 2.25 to 2.34. For the size factor (SMB), the loadings vary from 0.97 to 1.07. The SMB loadings are similar to loadings reported by Davis, Fama, and French (2000) and Fama and French (1995) for a portfolio of small public growth stocks. Davis et. al. find loadings on the size

[9] As discussed above, conditional on the company's valuation, the selection equation has a negative slope on RMRF. This translates to an upward-sloping selection boundary in Figure 2, mitigating the effect of dynamic selection on Beta.

factor ranging from 1.22 to 1.47. Fama and French report loadings between 0.99 and 1.44. For the book-to-market (HML) factor, we find negative loadings between -1.65 and -1.54. Davis, Fama, and French (2000) report loadings between -0.14 and -0.23, and Fama and French (1995) report loadings between -0.31 and -0.20, confirming that for venture capital backed private companies growth options represent a larger fraction of the total value than for publicly traded growth stocks. It is interesting that the size and book-to-market factors, which were developed to explain returns to publicly traded companies, appear to also capture patterns in the valuations of privately held entrepreneurial companies.

**** TABLE IV: THREE-FACTOR MODEL WITH SELECTION ****

C. Comparing Companies at Different Stages and Periods

Table V presents estimates with separate coefficient for investments in companies at different stages. We refer to four stages of development: "seed," "early," "late," and "mezzanine," as defined by Sahlman (1990). The table reveals interesting patterns. In all specifications, the intercept is largest for the seed stage, followed by the early and mezzanine stage, with the late stage having the lowest intercept. Seed investments have very little systematic risk, suggesting that, at this stage, most of the uncertainty is idiosyncratic. This is consistent with the definition of seed investments, which are primarily investments to develop young ideas or prototypes where the risk is mainly technological. The exposure to the market tends to increase with the stage of the investment. As the companies mature the option to become public companies becomes

more dominant in their valuations, increasing their exposure to market risks (however, see Berk, Green, and Naik (2004)).

The third specification includes the size and book-to-market factors. Again, the seed investments have no systematic exposure to any of the factors, but as the companies mature their exposures to the size factor range from 1.3 to 1.8. The HML factor has small insignificant loadings at the seed and mezzanine stages, but loadings around -1.8 for early stage investments increasing to loadings around -1.2 for late and mezzanine investments. Interpreting this exposure as a measure of growth options, it is consistent with the early stage having more rapid growth than the late and mezzanine stage investments. Interestingly, the measure of idiosyncratic volatility remains fairly constant across the four stages.

**** TABLE V: ESTIMATES BY COMPANY STAGE ****

In table VI, we report two specifications where the parameters in the valuation equation vary by time period. We see that the intercept is markedly lower in the post-2001 period, indicative of a recent lower performance of venture capital investments. Below, we find substantial differences in the alphas calculated for these three periods. The idiosyncratic risk is fairly constant across time periods, but the figures indicate that the very high betas predominant in the late nineties have abated in recent years.

**** TABLE VI: ESTIMATES BY INVESTMENT PERIOD****

D. Estimates with VC Factor

We define a separate VC factor by the monthly change in the logarithm of the total dollar volume of VC investments. This is motivated by Gompers and Lerner (2000) and Kaplan and Schoar (2005) who suggest that capital inflows into venture capital funds lead to higher valuations and subsequent poorer performance. This effect may introduce a risk factor that is specific to venture capital investments, and our factor is an attempt to provide a measure of the potential magnitude of this risk. In table VII, we see that the valuations load strongly on the VC factor, suggesting that there is substantial VC-specific risk that is not captured by the three-factor model. Including this factor reduces the loadings on the market factor substantially, from about 2.8 without the factor to about 1.0 with it. Similarly, the absolute magnitudes of the loadings on SMB and HML decline markedly with the factor. The positive coefficient in the selection equation confirms that the probability of observing a valuation increases with the aggregate amount of VC investments, not surprisingly.

One possible interpretation of these findings is that the new loadings on the market-, size-, and book-to-market-factors capture the "inherent" business risk of VC-backed companies, and the loading on the VC factor captures the "capital risk" arising from the effect of capital in- and out-flows on valuations. This interpretation, however, ignores the reverse causality of the valuations on capital flows, and the estimates may well overestimate the direct causal effect of capital flows on valuations. Nevertheless, the results are indicative of the potential magnitude of this effect, and suggest that there may be substantial VC-specific risk, possibly leaving even a diversified portfolio of VC

investments with substantially greater risk than predicted by standard models. Pricing this risk is difficult, however. It is unclear whether it is possible to construct a factor-mimicking portfolio of publicly traded stocks, making it difficult to assess the risk premium associated with this factor.

**** TABLE VII: ESTIMATES WITH VC FACTOR ****

V. Interpretation of Intercepts

Interpreting the economic magnitudes of the intercepts is not straightforward. For our estimates using log returns, the arithmetic alpha defined in equation (9) is calculated using the correction $\alpha = \delta + \frac{1}{2}\sigma^2 - \frac{1}{2}\beta(1-\beta)\sigma_m^2$. (The analogous adjustment for multi-factor specifications is in footnote 4.) One advantage of the Bayesian approach is that it is simple to compute the posterior distribution of alpha using these corrections even though the alpha is a non-linear function of the estimated parameters and σ^2 is not asymptotically Normal. The estimated alphas for the specifications in the previous tables are in Table VIII. We first report GLS and MCMC estimates without correcting for selection. The arithmetic GLS estimates provide a direct estimate of the alpha, whereas the alphas for the log-GLS and MCMC estimates are calculated using the correction. Without correcting for selection, we see high monthly alphas between 5.1% and 7.9%.

Correcting for selection, the alphas drop substantially, as indicated both in Table VIII and Figure 2. Figure 2 compares the posterior distributions of the alphas found using the MCMC estimates without selection correction from Table III and the estimates with

correction from the first specification in Table II. Correcting for selection, the market model specifications in Table II and the Fama-French specifications in Table IV show monthly alphas ranging from 3.3% to 3.5%. In the specifications that separate investments by the stage of the company, we find that seed investments have larger alphas and late-stage investments offer the lowest alphas. Finally, Table VIII shows substantial variation in the alphas over the three different time periods. In Figure 3 we plot the posterior distributions and see that the early period, from 1987-93, offered a moderate monthly alpha of around 1.6%, with zero alpha being within the simulated posterior distribution. This increased dramatically in the late nineties, during the dot-com boom, to a monthly alpha around 5.8%. The recent period 2001-2005 appears to have experienced more disappointing returns with average monthly alphas around -2.7%.

When interpreting the estimates of alpha as measures of risk-adjusted investment returns, it is important to keep in mind the specification of the model and the unit of analysis. Our estimates reflect the average monthly risk and return for companies receiving venture capital financing. Given a company and its current valuation, our estimates predict next month's valuation as a function of the market return and other observed variables. This is a natural starting point for understanding the risk and return properties of entrepreneurial companies, but it may not directly measure the investment returns earned by VCs or LPs, for several reasons: First, the investments are illiquid, cannot be traded, and appear to contain substantial systematic risk that is particular to VC investments. It is not clear how to adjust the return measures for the investors' inability to rebalance their portfolios and price the VC-specific risk. Second, our returns are gross

returns that do not account for the fees and carry paid by the LPs to the GPs. Third, the investments are not independent in the sense that is it not possible to participate in only some investments in a company without also participating in the other ones. Indeed, an important part of an early investment is that it provides a real option to invest in future rounds, should be company be successful. Fourth and probably most importantly, investors are concerned about the dollar-weighted return on their investments. Our estimates suggest that the highest returns are earned for seed stage investments, but the dollar amounts invested in these rounds are tiny compared to the early- and late-stage rounds. Computing the dollar-weighted returns would substantially complicate the algorithm. A simple back-of-the-envelope calculation provides a sense of the magnitude of this effect: We can weigh the company-stage alphas in table VIII by the percentage of dollars invested in each stage (from VentureSource). They report that 1% of VC dollars are invested in seed-stage companies, 45% and 50% are invested in early- and late-stage companies, respectively, leaving 4% for mezzanine rounds. With these figures, we calculate a simple dollar-weighted monthly alpha of about 2.5%.

**** TABLE VIII: ESTIMATES OF ALPHA ****

**** FIGURE 3: HISTOGRAMS OF ALPHA ****

VI. Conclusion

Empirical problems arise when estimating the risk and return of assets with infrequently observed valuations. We show that when the timing of the observed

valuations is endogenous, a dynamic sample selection problem can bias traditional measures of risk and return, and we introduce a new methodology to address this problem.

We estimate our model using data with venture capital investments in entrepreneurial companies, and our results suggest that the selection bias is substantial. Correcting for selection leads to substantially lower intercepts and higher estimates of risk exposures, both for systematic and idiosyncratic risk. These findings are robust across specifications of the pricing model and selection equation.

Our approach explicitly models the path of the unobserved valuations between the observed ones, accounting for the factor returns over this period and the fact that no valuation was observed during this interim period, which shifts down the conditional distribution of the valuations. From these valuations, we estimate various measures of risk exposures as well as the selection process. Due to the large number of unobserved valuation and selection variables, the model is numerically difficult to estimate. We present a Bayesian estimator, relying on insights from Gibbs sampling and Kalman Filtering, which is surprisingly tractable and robust given the complexity of the model.

Similar problems have been encountered in studies of real-estate indices and hedge fund performance, two other areas with infrequent and endogenous observations of valuations. Previous studies have struggled with addressing these problems adequately, but it may be possible to apply our methodology, with few modifications, to those areas as well.

Appendix A: Details of Estimation Procedure

For each company, the econometric model is given by equations (11) and (13) as

$$v(t) = v(t-1) + r + \delta + \beta\left(r_m(t) - r\right) + \varepsilon(t), \tag{19}$$

$$w(t) = Z'(t)\gamma_0 + v(t)\gamma_v + \eta(t), \tag{20}$$

with *i.i.d.* distributions $\varepsilon(t) \sim N(0, \sigma_v^2)$ and $\eta(t) \sim N(0,1)$. The augmented posterior distribution is simulated using a Gibbs sampler (Gelfand and Smith (1990)) with three blocks, containing the valuations, the selection variables, and the parameters, respectively, as described below. For simplicity, we suppress the dependence of the risk-free rate *r* on *t*.

A. Draw Latent Valuation Variables Using FFBS

The latent valuation variables for the interim period between two observed valuations are sampled conditional on the parameters, the selection variables, and the realized market (and other factor) returns. We use the Forward Filtering Backwards Sampling (FFBS) procedure (Carter and Kohn (1994) and Fruhwirth-Schnatter (1994)), which provides an efficient way to sample a path of state variables defined by a linear state space model. Since the error terms are assumed *i.i.d.* across firms, we can sample $v(t)$ separately for each firm.

Interpreting the econometric model as a linear state space model, $v(t)$ is the state variable, and the outcome equation (19) is the transition rule. Conditional on the

parameters, $r+\delta+\beta(r_m(t)-r)$ is an "observed" control acting on the state, and conditional on $w(t)$, the selection equation (20) is a noisy observation equation for the state.

The conditional distribution of the state vector of latent valuations is given by the identity (Lemma 2.1 in Carter and Kohn (1994))

$$p(v(1)\ldots v(T)\,|\,w^T)=p\{v(T)\,|\,w^T\}\prod_{t=1}^{T-1}p\{v(t)\,|\,w^t,v(t+1)\}, \tag{21}$$

where $w^t=\{w(1),\ldots,w(t)\}$ contains the selection variables up to time *t*. Simulating from this distribution requires two steps: a forward filtering and a backward sampling step. Define $m(t\,|\,j)=E\{v(t)\,|\,w^j\}$ and $s(t\,|\,j)=\mathrm{var}\{v(t)\,|\,w^j\}$ as the mean and variance of $v(t)$ conditional on the selection variables up to time *j*. Note that all conditional distributions are Normal and hence fully characterized by their means and variances (see Kalman (1960) and Anderson and Moore (1979)).

For the forward filtering step, for $t=1,\ldots,T$, we calculate $m(t\,|\,t)$ and $s(t\,|\,t)$ by iterating on the forward filter, through a forecasting and an updating part. The forecasting part involves the two equations

$$m(t+1\,|\,t)=m(t\,|\,t)+\left(r+\delta+\beta(r_m-r)\right), \tag{22}$$

and

$$s(t+1\,|\,t)=s(t\,|\,t)+\sigma^2\,. \tag{23}$$

For the updating part, as long as $v(t)$ remains unobserved, we update

$$m(t \mid t) = m(t \mid t-1) + K \cdot [w(t) - X'(t)\gamma_0 - m(t \mid t-1)\gamma_v], \tag{24}$$

where the Kalman gain K is given by

$$K = \frac{\gamma_v s(t \mid t-1)}{1 + \gamma_v^2 s(t \mid t-1)}. \tag{25}$$

When K is large, more weight is placed on the information from the selection equation. This happens when either γ_v or $s(t \mid t-1)$ is large, i.e. when either the selection equation is more informative about the valuations or when they are more uncertain. Further,

$$s(t \mid t) = s(t \mid t-1) \cdot (1 - \gamma_v K). \tag{26}$$

To estimate the model without correcting for selection, we force $\gamma_v = 0$. Then $m(t \mid t) = m(t \mid t-1)$ and $s(t \mid t) = s(t \mid t-1)$, and no information is used in periods where $v(t)$ is unobserved. In periods where $v(t)$ is observed, $m(t \mid t) = v_{OBS}(t)$ and $s(t \mid t) = 0$.

For the backward sampling part, $v(T)$ is first simulated from the Normal distribution with mean $m(T \mid T)$ and variance $s(T \mid T)$, as found above. For $t = T-1, \ldots, 1$, we simulate $v(t)$ from the conditional distribution $p\{v(t) \mid w^t, v(t+1)\}$. This distribution can be derived from a filtering problem where the draw of $v(t+1)$ provides an additional observation of $v(t)$. Hence, the distribution is

$$p\{v(t) \mid w^t, v(t+1)\} \sim N(r, q). \tag{27}$$

where

$$r = m(t \mid t) + G \cdot [v(t+1) - m(t+1 \mid t)]. \tag{28}$$

$$q = s(t \mid t) \cdot (1 - G). \tag{29}$$

for

$$G = \frac{s(t \mid t)}{s(t \mid t) + \sigma^2}. \tag{30}$$

B. Draw Selection Variables from Truncated Normal Distributions

The selection variables are sampled conditional on the valuations, parameters, and whether the valuation is observed or not. When the valuation is observed, the posterior distribution of the selection variable is

$$w(t) \mid Z, v, \gamma \sim LTN\left(Z'(t)\gamma_0 + v(t)\gamma_v, 1\right). \tag{31}$$

When it is unobserved, it is

$$w(t) \mid Z, v, \gamma \sim UTN\left(Z'(t)\gamma_0 + v(t)\gamma_v, 1\right). \tag{32}$$

Here, $LTN(\mu, \sigma^2)$ denotes a Normal distribution with mean μ and variance σ^2 truncated at zero from below, and *UTN* is the same distribution truncated at zero from above. Simulating this block is similar to simulating the (augmented) posterior distribution of a Probit model (Albert and Chib (1993)).

C. Draw Parameters Using a Bayesian Linear Regression

Conditional on $v(t)$ and $w(t)$, the distributions of δ, β, σ^2, and γ are given by two Bayesian linear regressions. Since $\varepsilon(t) \perp \eta(t)$ by assumption, we estimate the two equations separately.

In the valuation equation, δ, β, and σ^2 are defined by the regression of the excess returns $Y_v(t) = v(t) - v(t-1) - r$ (stacked for all companies) on a constant term and $r_m(t) - r$. Let $N(t)$ be the number of companies for which $v(t)$ exists, so $Y_v(t)$ is a $N(t) \times 1$ vector. Correspondingly, let $X_v(t)$ be a $N(t) \times 2$ matrix with a constant term and $r_m(t) - r$. Let Y_v and X_v contain $Y_v(t)$ and $X_v(t)$ stacked over time. The standard conjugate Normal-Inverse Gamma prior with prior parameters α_0, β_0, μ_0, and Σ_0 is

$$\sigma^2 \sim IG(\alpha_0, \beta_0) \tag{33}$$

$$\delta, \beta \mid \sigma^2 \sim N(\mu_0, \sigma^2 \Sigma_0^{-1}). \tag{34}$$

The posterior distributions for the parameters in the valuation equation are then (e.g. Rossi, Allenby, and McCulloch (2005)):

$$\sigma^2 \mid Y_v, X_v \sim IG(a, b) \tag{35}$$

$$\delta, \beta \mid \sigma^2, Y_v, X_v \sim N(\mu, \sigma^2 \Sigma^{-1}), \tag{36}$$

with parameters

$$a = a_0 + \sum_t N(t)\,, \tag{37}$$

$$b = b_0 + e'e + (\mu - \mu_0)'\Sigma_0(\mu - \mu_0)\,, \tag{38}$$

$$\Sigma = \Sigma_0 + X_v'X_v\,, \tag{39}$$

$$\mu = \Sigma^{-1}(\Sigma_0\mu_0 + X_v'Y_v)\,, \tag{40}$$

The vector *e* contains the stacked error terms $e = Y_v - X_v\mu$.

The selection equation is simpler. The parameters are given by a linear regression of $Y_s(t) = w(t)$ on $X_s(t) = [\underset{Z'(t)}{\quad} \quad v(t)]$, again stacked over companies. To identify the scale of the parameters, we normalize the variance of the error term to one. The prior distribution of $\gamma = [\gamma_0 \;\; \gamma_v]$ is

$$\gamma \sim N(\theta_0, \Omega_0^{-1})\,, \tag{41}$$

and the posterior distribution becomes

$$\gamma \mid Y_s, X_s \sim N(\theta, \Omega^{-1})\,, \tag{42}$$

with

$$\Omega = \Omega_0 + X_s'X_s\,, \tag{43}$$

$$\theta = \Omega^{-1}(\Omega_0\theta_0 + X_s'Y_s)\,. \tag{44}$$

D. Prior Distributions, Starting Values, and Miscellanea

We use diffuse priors for the parameters. We set the prior means of δ, β, γ_0 and γ_v, denoted μ_0 and θ_0 above, to zero. We set $\Sigma_0 = I/10,000$ and $\Omega_0 = I/100$, where I is the identity matrix. The prior distribution of σ^2 is an Inverse Gamma distribution with parameters $a_0 = 2.1$ and $b_0 = 1/600$, implying that $E(\sigma) = 4\%$ per month, and σ is between 1% and 12% (monthly) with 99% probability. Based on these choices, the priors for δ and β are $N(0,4^2)$, and the priors of γ_0 and γ_v are $N(0,10^2)$.[10] We start the Markov chain with δ, β and γ at zero and σ at 10%. We do not need starting values for $v(t)$ and $w(t)$, because $v(t)$ is the first variable we simulate and γ_v is zero initially, so our initial draws of $v(t)$ do not depend on $w(t)$.

Our implementation of the algorithm, using C++ and the GNU Scientific Library (GSL) on a 2.66 GHz Pentium 4 quad-core processor with 3.5Gb of RAM (using only a single core), takes about 30 minutes to simulate 6,000 draws of the Markov Chain.

[10] Our results are robust to using different priors. If we multiply the prior standard deviations on all parameters by 10, the base estimates of -0.0563 and 2.7510 change to -0.0496 and 2.6364 for the intercept and beta, respectively. Note that the more dispersed prior distributions lead to results closer to zero. The coefficients in the selection equation are less affected.

Appendix B: Robustness and Convergence

Below we confirm the robustness of our procedure: using simulated data, by testing for convergence, and relaxing the Normality assumption.

A. Estimation using Simulated Data

We simulate three sets of 1,000 datasets and estimate our model on those. For the first 1,000 datasets, we use Normal errors, as assumed in the model. For the second and third 1,000 datasets, we use Log-Normal and *t*-distributed errors to assess the robustness of the algorithm to misspecifications of the error distribution. As above, our procedure uses 1,000 iterations for burn-in and 5,000 iterations for the posterior distribution. We also compare the estimates to OLS, GLS and MCMC estimates without correcting for selection.

Each dataset contains 10 firms simulated over 120 periods. The valuation variables are simulated using equation (11), and the selection variables are simulated using equation (13), with the valuations being "observed" when $w(t) \geq 0$. The market return is assumed to be distributed *i.i.d*. Normal with mean zero and monthly standard deviation of $0.1/\sqrt{12}$. The results are reported in Tables B.I to B.III. These tables report the true parameters, the point estimates averaged over the 1,000 datasets, and the estimated standard error of the point estimates over the datasets is given in parentheses.[11] Overall, our algorithm seems to recover the underlying parameters very well, even with

[11] Calculated as the empirical standard error of the estimators for the 1,000 datasets divided by the square root of 1,000.

misspecified error distributions. The datasets used for the simulations are substantially smaller than the actual data, and the statistical power should be at least as good in the actual data as found here. As expected, the estimators that do not account for selection tend to underestimate the systematic risk. The GLS and MCMC estimators (without selection correction) produce very similar results and both overestimate the intercept and underestimate the volatility, consistent with the intuition behind the selection problem.

B. Convergence to Posterior Distribution

We use several tests to assess the convergence of the simulations to the posterior distribution: We plot the simulated parameters, their autocorrelation functions, and formally test for convergence using the Geweke (1992) and the Gelman and Rubin (1992) tests. These tests are all performed using the actual data and specification 1 in Table II. Convergence tests for the other specifications and the simulated data are similar.

Figure B.1 plots the parameter draws. They appear to converge quickly from their initial values to a stationary region of the parameter space. Most of the convergence appears within the first few hundred iterations, and there are no apparent subsequent drift or changes in the volatility.

To formally test for convergence, we first compute the Geweke (1992) convergence diagnostic. This diagnostic compares draws from the beginning and end of the chain (after discarding the initial 1,000 draws for burn-in). As suggested by Geweke (1992), we use a *Z*-score to test for equality of the means of the first 10% and the last 50% of the 5,000 remaining iterations, taking into account the auto-correlation of the

parameter draws using the Bartlett spectral density estimator of standard deviations. The results presented in Table B.IV show that we cannot reject equality of the means, suggesting that the subsamples are drawn from a stationary distribution and that our procedure has converged.

Our second convergence diagnostic uses the Gelman and Rubin (1992) potential scale reduction factor. This test is based on 10 chains each consisting of 1,000 burn-in iterations and 1,000 monitoring iterations, with starting values that are over-dispersed relative to the posterior distribution. If the chains have converged after the burn-in period, the variance within the chains should be similar to the variance between the chains. We draw starting values randomly as described in table B.IV, and calculate the *R*-statistic as the between-chain variance divided by the within-chain variance. Values of *R* above 1.1 are generally considered problematic. All our values are below 1.07, and we cannot formally reject the hypothesis that our chain has converged for any of the parameters.

C. Relaxing the Normality Assumption

One attractive feature of our procedure is that it preserves the Gibbs sampling and linear filtering properties when the distributional assumption is relaxed to mixtures of Normals (e.g. Carter and Kohn (1994)). Mixtures of Normals approximate a wide range of distributions, including skewed and fat-tailed distributions. We specify the density of the error term in the valuation equation as

$$f_{\varepsilon} = \sum_{i=1}^{K} p_i N(\mu_i, \sigma_i^2) .$$

This can be interpreted as if, with probability p_i, we draw $\varepsilon(t)$ from a Normal distribution with mean $N(\mu_i, \sigma_i^2)$. Note that only the combined mixture distribution, but not the individual underlying distributions, are identified, but this is not a problem for Bayesian estimators (see Rossi, Allenby, and McCulloch (2005) for details). We use prior distributions of $\sigma_i^2 \sim IG(2.1, 1/600)$ and $\mu_i \mid \sigma_i^2 \sim N(0, 100 \cdot \sigma_i^2)$. Figure B.3 presents parameter plots and the estimated parameters are in Table B.VI. In the top plot of Figure B.3 it is apparent that the individual mixtures are not identified and their means keep vacillating. However, the intercept is identified and converges quickly, as seen on the second plot. Similar patterns are observed for the variances and probabilities, as seen in the bottom two plots. In Table B.VI we see that the results are largely robust to relaxing the Normality assumption. The estimated mixtures show slightly positive skew and kurtosis, although the deviations from Normality are slight. The coefficients in the valuation and selection equations are largely unaffected. Given this evidence, the Normality assumption for the error term in the valuation equation seems unproblematic.

Bibliography

Ahn, Hyungtaik and James L. Powell, 1993, Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism, *Journal of Econometrics* 58, 3-29.

Albert, James and Siddhartha Chib, 1993, Bayesian Analysis of Binary and Polychotomous Response Data, *Journal of the American Statistical Association* 88, 669-679.

Amemiya, Takeshi, 1985. *Advanced Econometrics* (Harvard University Press, Cambridge, Massachusetts).

Anderson, Brian and John Moore, 1979. *Optimal Filtering* (Prentice Hall, New York).

Andrews, Donald W. K. and Marcia M.A. Schafgans, 1998, Semiparametric Estimation of the Intercept of a Sample Selection Model, *Review of Economics Studies* 65, 497-517.

Baquero, Guillermo, Jenke ter Horst, and Marno Verbeek, 2005, Survival, Look-Ahead Bias, and Persistence in Hedge Fund Performance, *Journal of Financial and Quatitative Analysis* 40, 493-517.

Berk, Jonathan, Richard Green, and Vasant Naik, 2004, Valuation and Return Dynamics of New Ventures, *Review of Financial Studies* 17, 1-35.

Campbell, John Y., Andrew W. Lo, and A. Craig MacKinlay, 1997. *The Econometrics of Financial Markets* (Princeton University Press, Princeton, New Jersey).

Carter, Chris K. and Robert J. Kohn, 1994, On Gibbs Sampling for State Space Models, *Biometrika* 81, 541-553.

Cochrane, John, 2005, The Risk and Return of Venture Capital, *Journal of Financial Economics* 75, 3-52.

Davis, James, Eugene Fama, and Kenneth French, 2000, Characteristics, Covariances, and Average Returns: 1929 to 1997, *Journal of Finance* 389-406.

Dimson, Elroy, 1979, Risk Measurement When Shares Are Subject to Infrequent Trading, *Journal of Financial Economics* 7, 197-226.

Driessen, Joost, Tse-Chun Lin, and Ludovic Phalippou, 2007, Measuring the Risk of Private Equity Funds: A New Approach, *working paper*.

Fama, Eugene and Kenneth French, 1995, Size and Book-to-Market Factors in Earnings and Returns, *Journal of Finance* 50, 131-155.

Fisher, Jeffrey, Dean H. Gatzlaff, David Geltner, and Donald R. Haurin, 2003, Controlling for the Impact of Variable Liquidity in Commercial Real Estate Price Indices, *Real Estate Economics* 31, 269-303.

Fruhwirth-Schnatter, Sylvia, 1994, Data Augmentation and Dynamic Linear Models, *Journal of Time Series Analysis* 15, 183-202.

Gatzlaff, Dean H. and Donald R. Haurin, 1997, Sample Selection Bias and Repeat-Sales Index Estimates, *Journal of Real Estate Finance and Economics* 14, 33-50.

Gelfand, Alan and Adrian Smith, 1990, Sampling Based Approaches to Calculating Marginal Densities, *Journal of the American Statistical Association* 85, 398-409.

Gelman, A. and Donald B. Rubin, 1992, Inference from Iterative Simulation Using Multiple Sequences, *Statistical Science* 7, 457-511.

Geman, Stuart and Donald Geman, 1984, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, 721-741.

Geweke, John, 1992. Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4* (Oxford University Press, Oxford).

Goetzmann, William N. and Liang Peng, 2006, Estimating House Price Indexes in the Presence of Seller Reservation Prices, *Review of Economics and Statistics* 88, 100-112.

Gompers, Paul and Josh Lerner, 1997, Risk and Reward in Private Equity Investments: The Challenge of Performance Assessment, *Journal of Private Equity* 5-12.

Gompers, Paul and Josh Lerner, 1999. *The Venture Capital Cycle* (MIT Press, Cambridge).

Gompers, Paul and Josh Lerner, 2000, Money Chasing Deals? The Impact of Fund Inflows on Private Equity Valuations, *Journal of Financial Economics* 55, 281-325.

Heckman, James, 1979, Sample Selection Bias as a Specification Error, *Econometrica* 47, 153-162.

Heckman, James, 1990, Varieties of Selection Bias, *American Economic Review* 80, 313-318.

ter Horst, Jenke and Marno Verbeek, 2007, Fund Liquidation, Self-Selection, and Look-Ahead Bias in the Hedge Fund Industry, *Review of Finance* 1-28.

Hwang, Min and John M. Quigley, 2003, Selectivity, Quality Adjustment and Mean Reversion in the Measurement of House Values, *Journal of Real Estate Finance and Economics* 28, 161-178.

Hwang, Min, John M. Quigley, and Susan E. Woodward, 2005, An Index for Venture Capital, 1987-2003, *Contributions to Economic Analysis & Policy* 4, 1-43.

Jagannathan, Ravi, Alexey Malakhov, and Dmitry Novikov, 2009, Do Hot Hands Exist among Hedge Fund Managers? An Empirical Evaluation, *Journal of Finance, forthcoming*.

Johannes, Michael and Nick Polson, 2006, MCMC Methods for Financial Econometrics, in Yacine Ait-Sahalia and Lars P. Hansen, eds.: *Handbook of Financial Econometrics, Forthcoming*.

Jones, Charles and Matthew Rhodes-Kropf, 2003, The Price of Diversifiable Risk in Venture Capital and Private Equity, *working paper*.

Kalman, Rudolf E., 1960, A New Approach to Linear Filtering and Prediction Problems, *Journal of Basic Engineering* 82, 35-45.

Kaplan, Steven and Antoinette Schoar, 2005, Private Equity Performance: Returns, Persistence, and Capital Flows, *Journal of Finance* 60, 1791-1823.

Kaplan, Steven, Berk Sensoy, and Per Strömberg, 2002, How Well Do Venture Capital Databases Reflect Actual Investments?, *working paper*.

Ljungqvist, Alexander and Matthew Richardson, 2003, The Cash Flow, Return and Risk Characteristics of Private Equity, *working paper*.

Peng, Liang, 2001, Building a Venture Capital Index, *working paper*.

Phalippou, Ludovic and Oliver Gottschalg, 2005, The Performance of Private Equity Funds, *Review of Financial Studies, forthcoming*.

Reyes, Jesse E., 1990, Industry Struggling to Forge Tools for Measuring Risk, *Venture Capital Journal* 30, 23-27.

Robert, Christian and George Casella, 2004. *Monte Carlo Statistical Methods* (Springer-Verlag, New York).

Rossi, Peter E., Greg Allenby, and Robert E. McCulloch, 2005. *Bayesian Statistics and Marketing* (John Wiley and Sons, Chichester, UK).

Sahlman, William, 1990, The Structure and Governance of Venture Capital Organizations, *Journal of Financial Economics* 27, 473-521.

Scholes, Myron S. and Joseph Williams, 1977, Estimating Betas from Nonsynchronous Data, *Journal of Financial Economics* 5, 309-328.

Tanner, Martin and Wing Wong, 1987, The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association* 82, 528-549.

Figure 1: Illustration of effect of selection on short and long-term observed average drift of a valuation process.

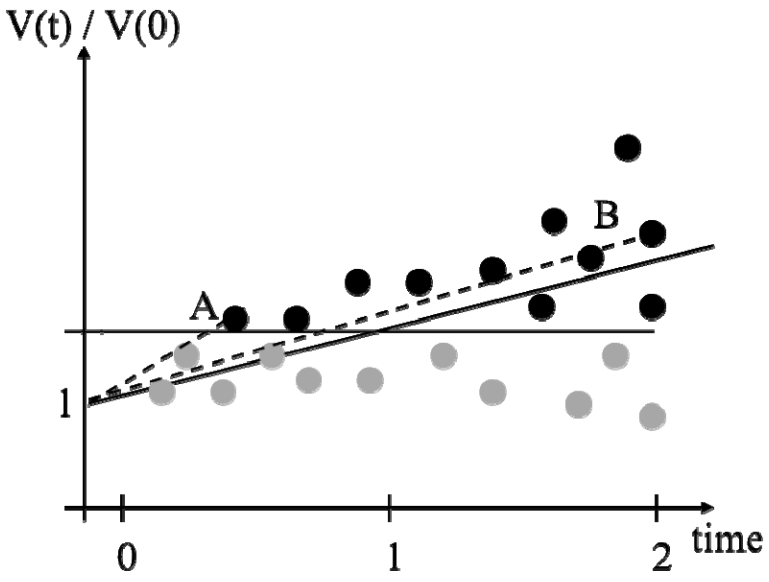


Figure 2: Illustration of selection bias on estimates of intercept, systematic risk, and idiosyncratic volatility.

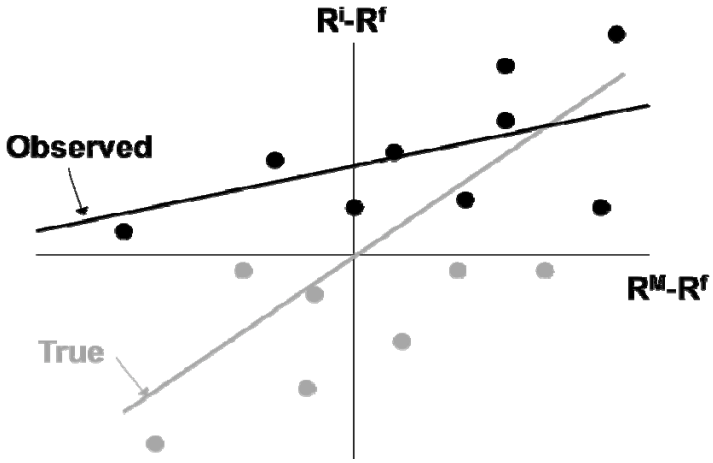
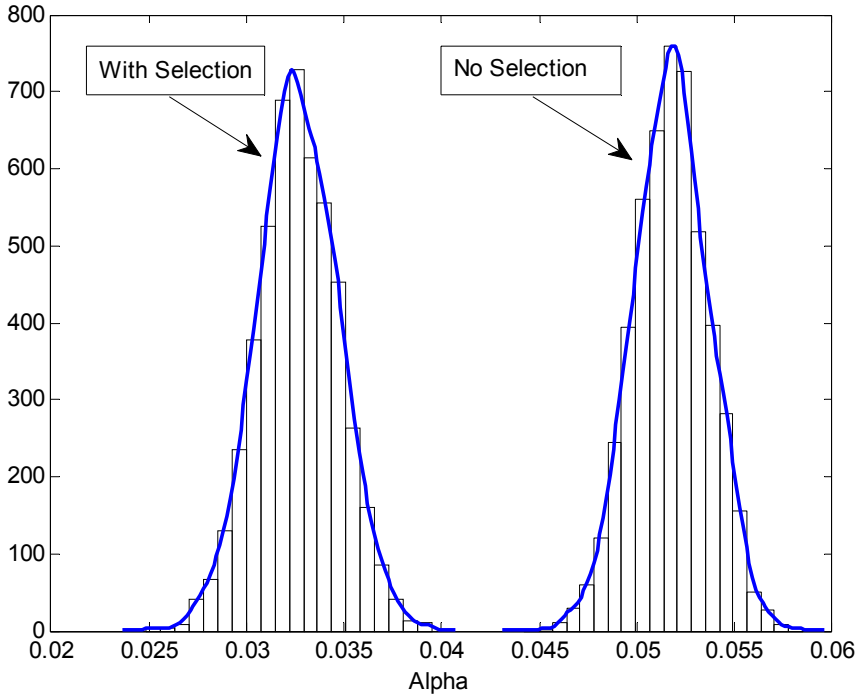


Figure 3: Posterior Distribution of the Monthly Excess Return: This figure plots the posterior distribution of monthly risk-adjusted excess returns, α, based on a one-factor market model in log-returns. In the top plot we estimate the model using an MCMC algorithm that uses the information in the selection equation to adjust for dynamic selection ("With Selection", table II model 1) and an MCMC algorithm that ignores the information in the selection equation ("No Selection", table III model 1). In the bottom plot we plot the distribution of α by sub-period, as described in table VI, using model 1.



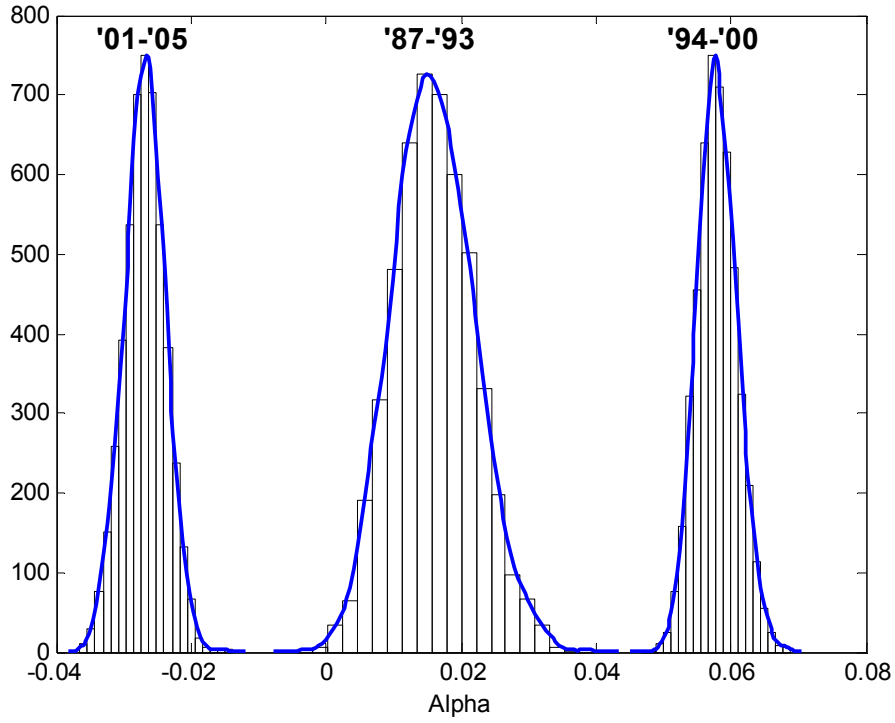


Table I: Descriptive Statistics: This table presents the number of rounds and companies in our sample relative to the full dataset, as well as the fraction of companies going public, being acquired, and liquidated.

	Data	Sample
Rounds	61,356	
Companies	18,237	1,934
Company Outcomes		
IPO	10.4%	10.3%
Acquisition	23.4%	23.3%
Liquidation	15.9%	23.0%
Unknown	50.4%	43.4%

Table II: Bayesian Estimates of One-Factor Market Model: The table presents MCMC estimates of the one-factor market model in monthly log returns, with selection correction. Factor and risk-free returns are from Kenneth French's website. All reported estimates are mean and standard deviations (in parentheses) of the simulated posterior distributions. In the valuation equation, Intercept and RMRF are the monthly intercept and the slope on the market log return (in excess of the risk-free rate). Sigma is the estimated monthly standard deviation of the error term. In the selection equation, Return is the log return earned since the previous financing event. Time is the time since this event (in years). Rounds contains the total number of VC investment rounds in the market in the same month (in 000s). ACQ and IPO contain the number of VC backed acquisitions and IPOs observed in this month (in 000s). The simulations use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. ***, **, and * denote whether zero is contained in the 1%, 5%, and 10% credible intervals, respectively.

	(1)	(2)	(3)	(4)
Valuation Equation				
Intercept	-0.0563 ***	-0.0566 ***	-0.0570 ***	-0.0571 ***
	(0.0016)	(0.0017)	(0.0015)	(0.0017)
RMRF	2.7510 ***	2.7900 ***	2.6773 ***	2.7013 ***
	(0.1127)	(0.1100)	(0.1071)	(0.1189)
Volatility	0.4109 ***	0.4119 ***	0.4135 ***	0.4131 ***
	(0.0050)	(0.0051)	(0.0045)	(0.0055)
Selection Equation				
Return	0.3321 ***	0.3368 ***	0.3502 ***	0.3508 ***
	(0.0079)	(0.0083)	(0.0097)	(0.0104)
Time	0.3666 ***	0.3777 ***	0.4139 ***	0.4137 ***
	(0.0202)	(0.0203)	(0.0211)	(0.0216)
Time Squared	-0.0361 ***	-0.0371 ***	-0.0405 ***	-0.0402 ***
	(0.0028)	(0.0028)	(0.0028)	(0.0027)
Acquisitions			6.9829 ***	6.4927 ***
			(1.0674)	(1.0446)
IPOs			-1.8940 *	-1.5898
			(1.0304)	(1.0137)
Rounds			0.3083 ***	0.3267 ***
			(0.0788)	(0.0793)
RMRF		-0.7095 ***		-0.4747 ***
		(0.1653)		(0.1656)
Constant	-1.9290 ***	-1.9331 ***	-2.2637 ***	-2.2588 ***
	(0.0170)	(0.0162)	(0.0275)	(0.0275)

Table III: OLS, GLS, and MCMC Estimates: The table presents OLS, GLS and MCMC estimates of the market model and the Fama-French 3-factor model in monthly log returns, without selection correction. Factor and risk-free returns are from Kenneth French's website. The OLS estimator regresses the log returns to the companies on the factor log returns. The GLS estimator scales each observation with the inverse of the square-root of the time since last financing round. MCMC estimates are the mean and standard deviation of the parameters' simulated posterior distribution, without correcting for selection bias (i.e. forcing $\gamma_v = 0$ in the estimation). RMRF is the return on the market in excess of risk-free rate, SMB is the small-minus-big portfolio, and HML the high-minus-low book-to-market portfolio. For the GLS and MCMC estimators, Sigma is the estimated monthly standard deviation of the error term from this regression. For the OLS estimator, Sigma is the estimated standard deviation of the error term (it does not have a time interpretation). The MCMC estimator use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively, or, for Bayesian estimates, whether zero is contained in the 1%, 5%, and 10% credible intervals, respectively.

Panel A: OLS

	(1)		(2)	
	Coef.	Std. Err.	Coef.	Std. Err.
Intercept	-0.0286	(0.0013) ***	-0.0221	(0.0016) ***
RMRF	2.0766	(0.1003) ***	1.8104	(0.1130) ***
SMB			-0.3258	(0.1710) *
HML			-1.0429	(0.1390) ***
Sigma	1.3695		1.3536	

Panel B: GLS

	(1)		(2)	
	Coef.	Std. Err.	Coef.	Std. Err.
Intercept	-0.0167	(0.0019) ***	-0.0110	(0.0021)
RMRF	2.2906	(0.1166) ***	2.1012	(0.1256) ***
SMB			-0.3581	(0.1915) *
HML			-0.9726	(0.1512) ***
Sigma	0.4156		0.4117	

Panel C: MCMC

	(1)		(2)	
	Mean	Std. Dev.	Mean	Std. Dev.
Intercept	-0.0159	(0.0015) ***	-0.0115	(0.0017) ***
RMRF	2.6624	(0.1170) ***	2.2631	(0.1145) ***
SMB			1.1377	(0.1747) ***
HML			-1.2435	(0.1340) ***
Sigma	0.3566	(0.0036) ***	0.3509	(0.0037) ***

Table IV: Estimates of Fama-French Three-Factor Model with Selection Correction: The table presents the posterior means and standard deviations (in brackets) of the parameters of the Fama-French model in log returns, with selection correction. Factor and risk-free returns are from Kenneth French's website. RMRF is the return on the market in excess of risk-free rate, SMB is the small-minus-big portfolio, and HML the high-minus-low book-to-market portfolio. Sigma is the monthly standard deviation of the error term. In the selection equation, Return is the log-return earned since the previous financing event, and Time is the time since this event (in years). Rounds measures the total number of VC investment rounds in the market in the same month (in 000s). Similarly, Acquisitions and IPOs contain the number of VC backed acquisitions and IPOs observed in this month (in 000s). MCMC simulations use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. ***, **, and * denote whether zero is contained in the 1%, 5%, and 10% credible intervals, respectively.

	(1)	(2)	(3)	(4)
Valuation Equation				
Intercept	-0.0538 ***	-0.0539 ***	-0.0544 ***	-0.0548 ***
	(0.0018)	(0.0018)	(0.0018)	(0.0019)
RMRF	2.2972 ***	2.3430 ***	2.2532 ***	2.3048 ***
	(0.1140)	(0.1090)	(0.1203)	(0.1208)
SMB	1.0651 ***	1.0168 ***	0.9728 ***	0.9759 ***
	(0.1608)	(0.1782)	(0.1790)	(0.1807)
HML	-1.6391 ***	-1.6513 ***	-1.5425 ***	-1.5487 ***
	(0.1258)	(0.1290)	(0.1339)	(0.1329)
Sigma	0.4033 ***	0.4038 ***	0.4048 ***	0.4060 ***
	(0.0050)	(0.0040)	(0.0044)	(0.0053)
Selection Equation				
Return	0.3311 ***	0.3374 ***	0.3462 ***	0.3509 ***
	(0.0094)	(0.0091)	(0.0089)	(0.0105)
Time	0.3673 ***	0.3752 ***	0.4067 ***	0.4115 ***
	(0.0212)	(0.0217)	(0.0207)	(0.0218)
Time Squared	-0.0362 ***	-0.0367 ***	-0.0398 ***	-0.0399 ***
	(0.0029)	(0.0029)	(0.0029)	(0.0029)
Acquisitions			6.9160 ***	6.3833 ***
			(1.0406)	(1.1412)
IPOs			-1.8341 *	-1.7884 *
			(0.9853)	(1.0304)
Rounds			0.2746 ***	0.3043 ***
			(0.0765)	(0.0833)
RMRF		-0.6025 ***		-0.3886 **
		(0.1670)		(0.1659)
SMB		0.0682		-0.1002
		(0.2296)		(0.2336)
HML		0.7097 ***		0.6203 ***
		(0.1903)		(0.2094)
Constant	-1.9340 ***	-1.9370 ***	-2.2488 ***	-2.2481 ***
	(0.0169)	(0.0166)	(0.0270)	(0.0273)

Table V: Estimates by Stage of Development of Entrepreneurial Company: The table presents the posterior means and standard deviations (in brackets) of the parameters of the market model and Fama-French model in log returns, by company stage. Factors and risk-free returns are from Kenneth French's website. RMRF is the return on the market in excess of risk-free rate, SMB is the small-minus-big portfolio, and HML the high-minus-low book-to-market portfolio. Sigma is the monthly standard deviation of the error term. In the selection equation, Return is the log-return earned since the previous financing event, and Time is the time since this event (in years). Rounds measures the total number of VC investment rounds in the market in the same month (in 000s). Similarly, Acquisitions and IPOs contain the number of VC backed acquisitions and IPOs observed in this month (in 000s). The simulations use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. ***, **, and * denote whether zero is contained in the 1%, 5%, and 10% credible intervals, respectively.

	(1)			(2)			(3)			(4)		
	Mean	Std.Dev.		Mean	Std.Dev.		Mean	Std.Dev.		Mean	Std.Dev.	
Valuation Equation												
Intercept												
seed	0.0436	(0.0108)	***	0.0452	(0.0104)	***	0.0434	(0.0106)	***	0.0461	(0.0112)	***
early	-0.0398	(0.0020)	***	-0.0405	(0.0020)	***	-0.0391	(0.0021)	***	-0.0397	(0.0022)	***
late	-0.0920	(0.0031)	***	-0.0922	(0.0033)	***	-0.0894	(0.0036)	***	-0.0892	(0.0036)	***
mezz	-0.0517	(0.0130)	***	-0.0516	(0.0130)	***	-0.0609	(0.0153)	***	-0.0630	(0.0143)	***
RMRF												
seed	0.7414	(0.7914)		0.5827	(0.7556)		0.7270	(0.7254)		0.4688	(0.8176)	
early	2.7425	(0.1267)	***	2.6633	(0.1309)	***	2.1774	(0.1317)	***	2.1693	(0.1424)	***
late	2.6281	(0.2210)	***	2.5053	(0.1877)	***	2.3840	(0.2204)	***	2.3481	(0.2319)	***
mezz	5.8885	(0.9108)	***	5.5939	(0.9100)	***	5.3149	(1.0087)	***	5.0712	(0.9047)	***
SMB												
seed							-0.1013	(0.6441)		-0.1443	(0.6081)	
early							1.4233	(0.2200)	***	1.3245	(0.2202)	***
late							0.5772	(0.3924)		0.4167	(0.3982)	
mezz							1.7806	(1.1654)		1.8336	(1.0111)	*
HML												
seed							0.5291	(0.4820)		0.5165	(0.5019)	
early							-1.8732	(0.1520)	***	-1.7795	(0.1542)	***
late							-1.2142	(0.1632)	***	-1.0380	(0.2679)	***
mezz							-1.2195	(0.9704)		-1.0938	(0.9146)	
Sigma												
seed	0.3434	(0.0155)	***	0.3417	(0.0149)	***	0.3415	(0.0168)	***	0.3404	(0.0151)	***
early	0.3880	(0.0056)	***	0.3886	(0.0051)	***	0.3784	(0.0053)	***	0.3800	(0.0054)	***
late	0.4396	(0.0100)	***	0.4386	(0.0108)	***	0.4397	(0.0093)	***	0.4392	(0.0109)	***
mezz	0.3930	(0.0350)	***	0.3810	(0.0314)	***	0.3761	(0.0332)	***	0.3664	(0.0331)	***
Selection Equation												
Return	0.3339	(0.0094)	***	0.3463	(0.0102)	***	0.3344	(0.0094)	***	0.3466	(0.0095)	***
Time	0.3738	(0.0223)	***	0.4089	(0.0202)	***	0.3785	(0.0210)	***	0.4092	(0.0225)	***
Time Sq	-0.0353	(0.0029)	***	-0.0386	(0.0027)	***	-0.0360	(0.0029)	***	-0.0386	(0.0029)	***
Acquisitions				6.6045	(1.0821)	***				6.5016	(1.0782)	***
IPOs				-1.5986	(0.9941)					-1.7514	(0.9935)	*
Rounds				0.3142	(0.0789)	***				0.2876	(0.0786)	***
RMRF	-0.6848	(0.1691)	***	-0.4554	(0.1748)	***	-0.5955	(0.1639)	***	-0.3376	(0.1808)	*
SMB							-0.0739	(0.2232)		-0.0936	(0.2428)	
HML							0.6960	(0.1804)	***	0.6054	(0.2028)	***
Constant	-1.9364	(0.0167)	***	-2.2589	(0.0263)	***	-1.9433	(0.0172)	***	-2.2488	(0.0279)	***

Table VI: Estimates by Investment Period: The table presents the posterior means and standard deviations (in brackets) of the parameters of the market model in log returns, by time period. RMRF is the return on the market in excess of risk-free rate (in logs, from Ken French's website). Sigma is the monthly standard deviation of the error term. In the selection equation, Return is the log-return earned since the previous financing event, and Time is the time since this event (in years). The simulations use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. ***, **, and * denote whether zero is contained in the 1%, 5%, and 10% credible intervals, respectively.

	(1)		(2)	
	Mean	Std.Dev.	Mean	Std.Dev.
Valuation Equation				
Intercept				
87-'93	-0.0387	(0.0055) ***	-0.0399	(0.0057) ***
94-'00	-0.0332	(0.0029) ***	-0.0341	(0.0029) ***
01-'05	-0.0926	(0.0029) ***	-0.0932	(0.0032) ***
RMRF				
87-'93	0.3814	(0.6710)	0.5015	(0.6245)
94-'00	2.5005	(0.2047) ***	2.5582	(0.1934) ***
01-'05	1.0855	(0.1837) ***	1.0554	(0.1745) ***
Sigma				
87-'93	0.3296	(0.0118) ***	0.3316	(0.0116) ***
94-'00	0.4185	(0.0053) ***	0.4192	(0.0059) ***
01-'05	0.3622	(0.0088) ***	0.3664	(0.0091) ***
Selection Equation				
Return	0.3348	(0.0083) ***	0.3393	(0.0089) ***
Time	0.3705	(0.0195) ***	0.3794	(0.0199) ***
Time Squared	-0.0358	(0.0027) ***	-0.0366	(0.0028) ***
RMRF			-0.4644	(0.1667) ***
Constant	-1.9391	(0.0161) ***	-1.9415	(0.0152) ***

Table VII: Estimates with VC Factor: The table presents posterior means and standard deviations (in brackets) of the parameters of the market model and the Fama-French 3-factor model in log returns, augmented with a VC specific factor. Factors and risk-free returns are from Kenneth French's website. RMRF is the return on the market in excess of the risk-free rate, SMB is the small-minus-big portfolio, and HML the high-minus-low book-to-market portfolio. VC Factor is the log-change in Dollars, where Dollars measures the total dollar volume of VC investments in a particular month. Sigma is the monthly standard deviation of the error term. In the selection equation, Return is the log-return earned since the previous financing event, and Time is the time since this event (in years). Rounds measures the total number of VC investment rounds in the market in the same month (in 000s). Similarly, Acquisitions and IPOs contain the number of VC backed acquisitions and IPOs observed in this month (in 000s). The simulations use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. ***, **, and * denote whether zero is contained in the 1%, 5%, and 10% credible intervals, respectively.

	(1)	(2)	(3)	(4)
Valuation Equation				
Intercept	-0.0537 ***	-0.0540 ***	-0.0527 ***	-0.0525 ***
	(0.0016)	(0.0016)	(0.0018)	(0.0019)
RMRF	0.9345 ***	1.0659 ***	0.9791 ***	1.1644 ***
	(0.1488)	(0.1713)	(0.1555)	(0.1756)
SMB			0.5201 ***	0.5435 ***
			(0.1915)	(0.1739)
HML			-1.0093 ***	-1.0556 ***
			(0.1290)	(0.1215)
VC Factor	0.5816 ***	0.5460 ***	0.4773 ***	0.4289 ***
	(0.0369)	(0.0377)	(0.0394)	(0.0411)
Sigma	0.4048 ***	0.4048 ***	0.4035 ***	0.4014 ***
	(0.0053)	(0.0045)	(0.0048)	(0.0045)
Selection Equation				
Return	0.3567 ***	0.3546 ***	0.3561 ***	0.3560 ***
	(0.0094)	(0.0089)	(0.0091)	(0.0104)
Time	0.4091 ***	0.4038 ***	0.4146 ***	0.4064 ***
	(0.0207)	(0.0209)	(0.0216)	(0.0243)
Time Squared	-0.0396 ***	-0.0387 ***	-0.0400 ***	-0.0390 ***
	(0.0028)	(0.0027)	(0.0030)	(0.0031)
Acquisitions	7.3768 ***	7.6483 ***	7.2524 ***	7.3702 **
	(1.0357)	(1.1640)	(1.0105)	(1.1041)
IPOs	-3.1900 ***	-3.1766 ***	-3.1626 ***	-3.0994 ***
	(1.0098)	(1.0451)	(0.9949)	(1.0286)
Rounds	0.2134 ***	0.1791 **	0.2251 ***	0.1866 **
	(0.0771)	(0.0849)	(0.0756)	(0.0807)
RMRF		-0.3309 *		-0.2997 *
		(0.1697)		(0.1827)
SMB				-0.1836
				(0.2344)
HML				0.4435 **
				(0.1991)
VC Factor		0.0838 ***		0.0950 ***
		(0.0274)		(0.0275)
Constant	-2.2136 ***	-2.2071 ***	-2.2207 ***	-2.2067 ***
	(0.0264)	(0.0289)	(0.0267)	(0.0294)

Table VIII: Monthly Risk-Adjusted Excess Returns: The table presents means, standard deviations, and percentiles of the posterior distributions of the monthly risk-adjusted excess returns (alphas).

	mean	std.dev.	1	5	50	95	99
Table III: No selection							
Model 1							
GLS	0.0512	(arithmetic)					
GLS	0.0681	(log)					
MCMC	0.0517	(0.0019)	0.0472	0.0486	0.0517	0.0549	0.0562
Model 2							
GLS	0.0574	(arithmetic)					
GLS	0.0794	(log)					
MCMC	0.0560	(0.0022)	0.0513	0.0525	0.0559	0.0598	0.0613
Table II: One-factor market model							
Model 1	0.0326	(0.0021)	0.0277	0.0292	0.0326	0.0361	0.0375
Model 2	0.0327	(0.0020)	0.0281	0.0294	0.0326	0.0361	0.0377
Model 3	0.0329	(0.0021)	0.0283	0.0296	0.0329	0.0365	0.0380
Model 4	0.0325	(0.0021)	0.0274	0.0290	0.0325	0.0361	0.0376
Table IV: Fama-French three-factor model							
Model 1	0.0351	(0.0023)	0.0299	0.0313	0.0351	0.0390	0.0405
Model 2	0.0355	(0.0023)	0.0300	0.0317	0.0355	0.0393	0.0407
Model 3	0.0345	(0.0022)	0.0297	0.0311	0.0344	0.0383	0.0398
Model 4	0.0349	(0.0024)	0.0294	0.0310	0.0349	0.0389	0.0405
Table V: By stage							
Model 1							
seed	0.1031	(0.0117)	0.0781	0.0850	0.1026	0.1235	0.1325
early	0.0400	(0.0024)	0.0346	0.0362	0.0399	0.0440	0.0462
late	0.0087	(0.0038)	-0.0002	0.0023	0.0088	0.0148	0.0173
mezz	0.0528	(0.0196)	0.0129	0.0233	0.0512	0.0881	0.1051
Model 2							
seed	0.1040	(0.0112)	0.0801	0.0867	0.1034	0.1233	0.1325
early	0.0392	(0.0023)	0.0340	0.0355	0.0392	0.0430	0.0445
late	0.0081	(0.0039)	-0.0004	0.0019	0.0079	0.0148	0.0174
mezz	0.0464	(0.0185)	0.0084	0.0180	0.0453	0.0785	0.0947
Model 3							
seed	0.1025	(0.0120)	0.0777	0.0845	0.1018	0.1223	0.1385
early	0.0408	(0.0026)	0.0348	0.0366	0.0408	0.0452	0.0470
late	0.0137	(0.0048)	0.0027	0.0057	0.0138	0.0217	0.0247
mezz	0.0399	(0.0212)	-0.0017	0.0085	0.0380	0.0777	0.0973
Model 4							
seed	0.1049	(0.0126)	0.0779	0.0849	0.1042	0.1266	0.1376
early	0.0404	(0.0027)	0.0342	0.0360	0.0403	0.0448	0.0469
late	0.0131	(0.0051)	0.0025	0.0051	0.0128	0.0217	0.0261
mezz	0.0311	(0.0196)	-0.0070	0.0014	0.0293	0.0667	0.0826
Table VI: By time period							
Model 1							
'87-'93	0.0159	(0.0060)	0.0028	0.0064	0.0157	0.0261	0.0306
'94-'00	0.0580	(0.0030)	0.0515	0.0533	0.0579	0.0631	0.0653
'01-'05	-0.0269	(0.0031)	-0.0339	-0.0320	-0.0268	-0.0218	-0.0198
Model 2							
'87-'93	0.0153	(0.0055)	0.0034	0.0064	0.0151	0.0246	0.0286
'94-'00	0.0576	(0.0030)	0.0506	0.0527	0.0576	0.0625	0.0646
'01-'05	-0.0259	(0.0031)	-0.0331	-0.0311	-0.0260	-0.0209	-0.0181
Table B.VI: Mixtures of Normals							
K = 2	0.0395	(0.0024)	0.0339	0.0356	0.0395	0.0434	0.0448
K = 3	0.0384	(0.0030)	0.0311	0.0332	0.0386	0.0431	0.0448

Figure B.1: Trace Plots: Plots of the parameter draws from the MCMC estimation of the one-factor market model in monthly log returns, with selection correction (model 1 in table II). In the valuation equation, δ and β are the monthly intercept and the slope on the market log return (in excess of the risk-free rate). σ is the estimated monthly standard deviation of the error term. In the selection equation, γ_0 is the loading on the intercept, γ_1 is the loading on the log return since the previous financing event, and γ_2 and γ_3 are loadings on the time since the last financing event (in years), and its squared value.

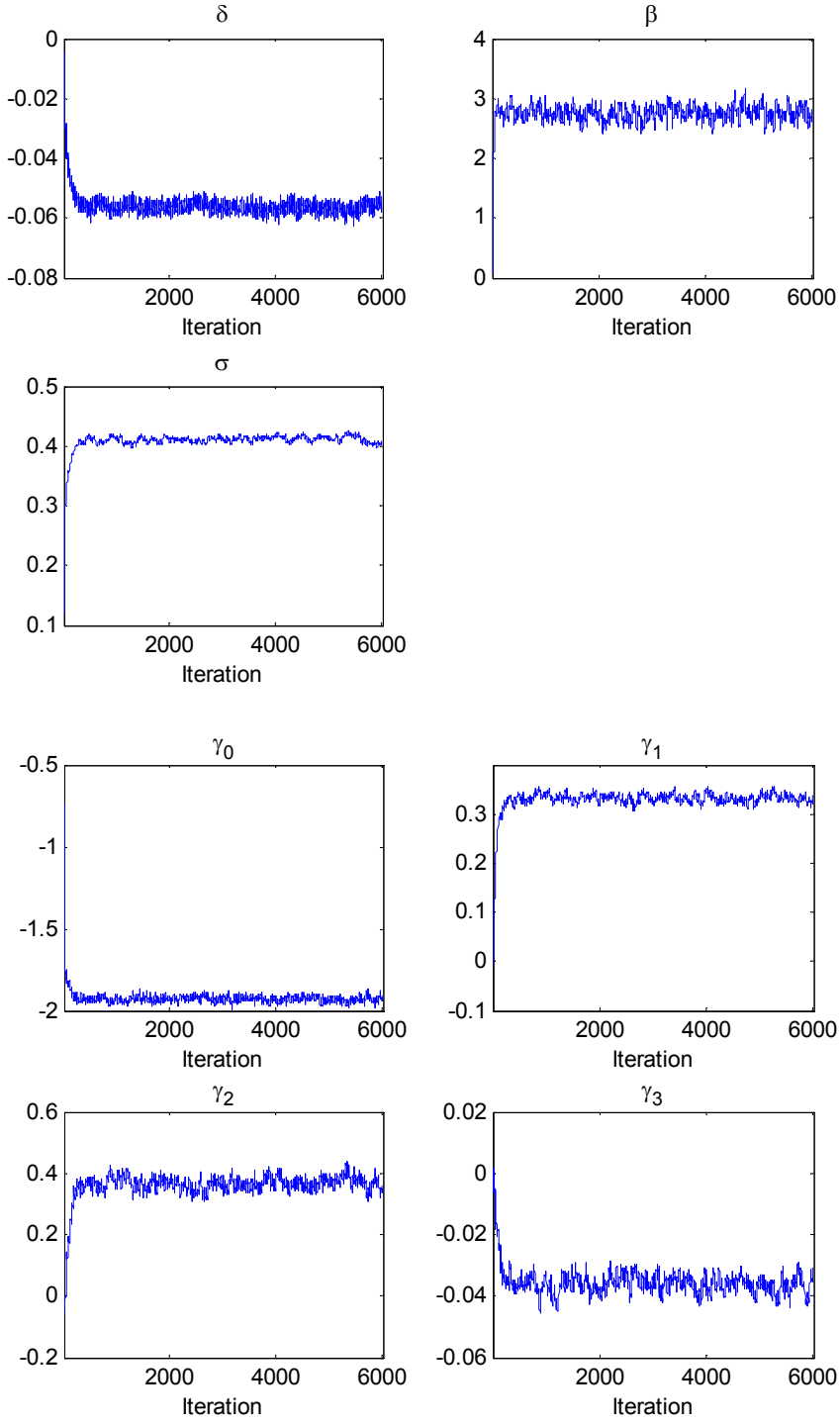


Figure B.2: Auto-correlation Functions: Plots of the autocorrelation functions of the MCMC draws of the one-factor market model in monthly log returns, with selection correction (model 1 in table II). The autocorrelations are calculated from 5,000 iterations of the MCMC algorithm, after discarding the first 1,000 draws. In the valuation equation, δ and β are the monthly intercept and the slope on the market log return (in excess of the risk-free rate). σ is the estimated monthly standard deviation of the error term. In the selection equation, γ_0 is the loading on the intercept, γ_1 is the loading on the log return since the previous financing event, and γ_2 and γ_3 are loadings on the time since the last financing event (in years), and its squared value.

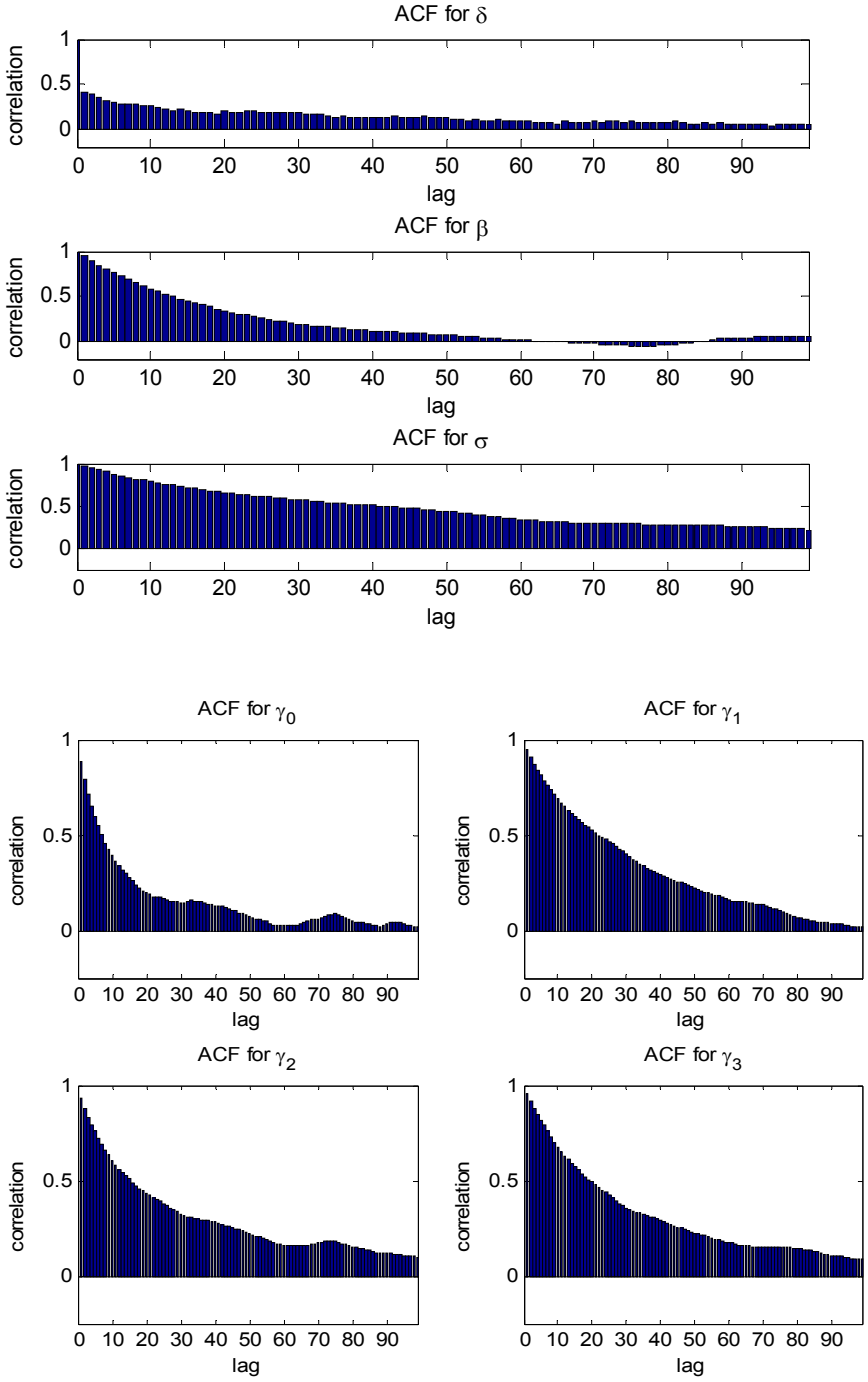


Figure B.3: Trace Plots of Mixture of 2 Normals Error Term Parameters: Plots of parameter draws of the MCMC estimation of the mixture model with 2 Normals in the error term, as described in table B.VI. We graph the individual error mixture distribution means (μ_1 and μ_2), and the valuation equation intercept (ν) in the top plot. In the second plot we show $\delta = \nu + \sum_{i=1}^{K} p_i \mu_i$, which represents the intercept in the valuation equation when the error term has mean zero. The standard deviations of the mixture distributions are σ_1 and σ_2 are in the third plot, and the last plot shows the probabilities of the mixture distributions.

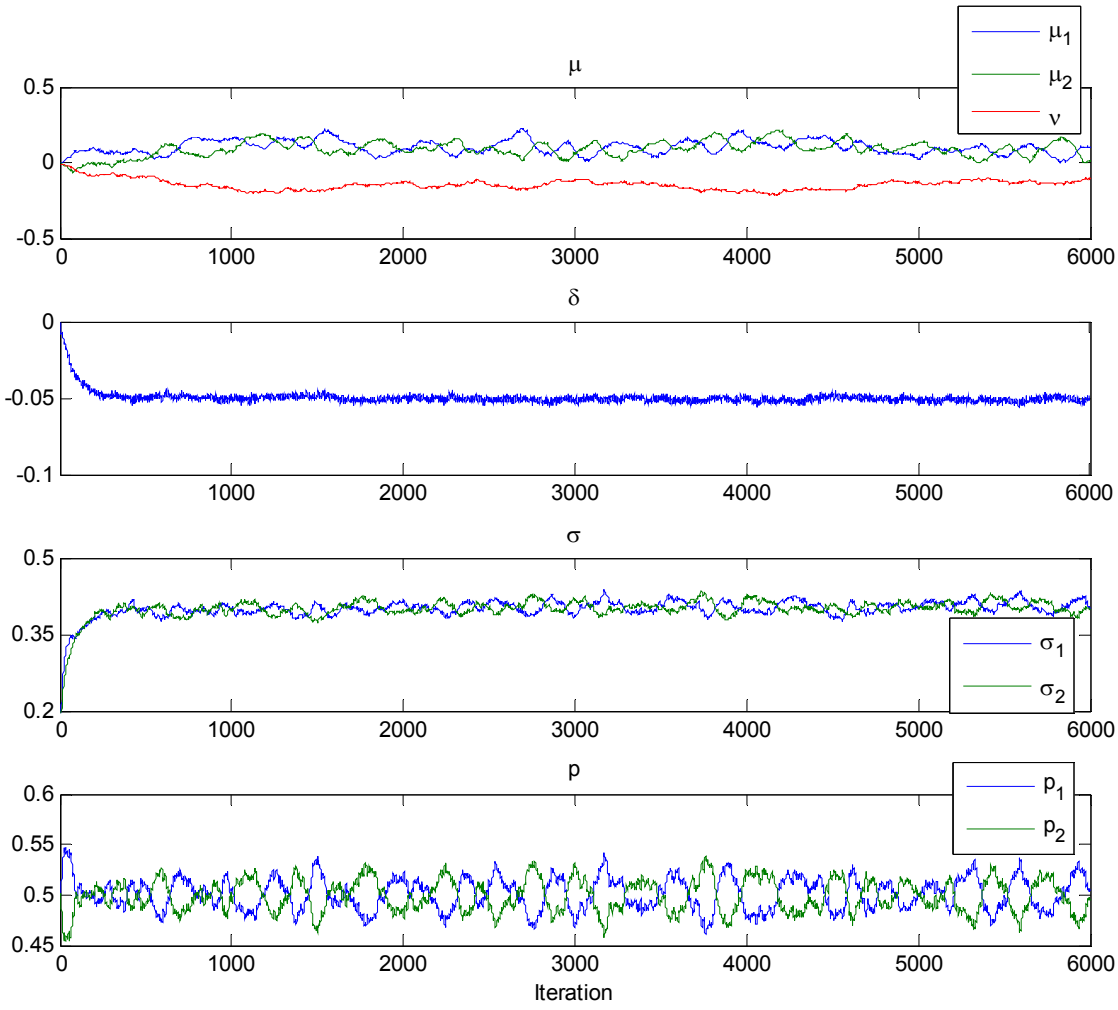


Table B.I: Estimates using Simulated Data Estimation results from 1,000 simulated datasets of 10 firms over 120 months. The simulated model is

$$v(t) = v(t-1) + r + \delta + \beta(r_m(t) - r) + \varepsilon(t)$$
$$w(t) = \gamma_0 + \gamma_1 v(t) + \gamma_2 \tau + \gamma_3 \tau^2 + \eta(t)$$

where $v(t) = \ln(V(t))$ is observed when the latent selection variable $w(t) \geq 0$. The log-market return $r_m(t)$ is drawn from an i.i.d. $N(0, 0.1^2/12)$, and τ is the time since the last observed valuation. The error terms $\varepsilon(t) \sim N(0, \sigma^2)$ and $\eta(t) \sim N(0,1)$ are independent of each other. We set the risk-free rate r to zero. Other parameter values used to simulate the model are shown in the column labeled "True". The OLS and GLS methods are explained in the main text. The MCMC (no selection) method forces $\gamma_v = 0$, as described in the paper. The MCMC (w/ selection) method is our dynamic selection algorithm detailed in appendix A. Both MCMC methods use the same priors as in Appendix A. For each variable, the first number is the mean of the point estimates across datasets (posterior means for MCMC results). The number in parentheses is the standard error of the estimates across datasets.

	True	OLS	GLS	MCMC (no selection)	MCMC (w/ selection)
Valuation Equation					
Intercept	0.0	-0.0038	0.0077	0.0077	0.0001
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
RMRF	3.0	1.1926	2.3578	2.3585	3.0100
		(0.0122)	(0.0123)	(0.0124)	(0.0118)
Sigma	0.1	0.1468	0.0875	0.0864	0.0990
		(0.0003)	(0.0002)	(0.0002)	(0.0003)
Selection Equation					
Return	10.0				10.6303
					(0.0484)
Time	0.1				0.1078
					(0.0011)
Time-Squared	0.0				0.0000
					(0.0000)
Constant	-1.0				-1.0241
					(0.0052)

Table B.II: Estimates using Simulated Data with *t*-distributed Errors: Simulations of the model described in table B.I, but with $\varepsilon(t) \sim 0.0775 \cdot t_5$, where t_5 is the Student t-distribution with 5 degrees of freedom. The distribution of $\varepsilon(t)$ is symmetric with mean zero and standard deviation 0.1, and excess kurtosis (in excess of the Normal distribution) of 6. We refer the reader to table B.I for more details.

	True	OLS	GLS	MCMC (no selection)	MCMC (w/ selection)
Valuation Equation					
Intercept	0.0	-0.0037	0.0077	0.0077	0.0002
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
RMRF	3.0	1.1774	2.3313	2.3338	3.0632
		(0.0132)	(0.0123)	(0.0123)	(0.0143)
Sigma	0.1	0.1512	0.0901	0.0889	0.1030
		(0.0004)	(0.0003)	(0.0003)	(0.0004)
Selection Equation					
Return	10.0				10.2325
					(0.0538)
Time	0.1				0.1048
					(0.0011)
Time-Squared	0.0				-0.0001
					(0.0000)
Constant	-1.0				-0.9943
					(0.0051)

Table B.III: Estimates using Simulated Data with Log-Normal Errors: Simulations of the model described in table B.I, but with $\varepsilon(t) \sim LN(0, 0.0993^2) - \exp(0.0993^2/2)$. The distribution of $\varepsilon(t)$ has mean zero and standard deviation 0.1, skewness of 1.8346 and excess kurtosis (in excess of the Normal distribution) of 0.16. We refer the reader to table B.I for more details.

	True	OLS	GLS	MCMC (no selection)	MCMC (w/ selection)
Valuation Equation					
δ	0.0	-0.0038	0.0078	0.0078	0.0002
		(0.0001)	(0.0001)	(0.0001)	(0.0001)
β	3.0	1.1556	2.3050	2.3054	3.0257
		(0.0132)	(0.0132)	(0.0133)	(0.0120)
σ	0.1	0.1510	0.0913	0.0901	0.1037
		(0.0003)	(0.0002)	(0.0002)	(0.0003)
Selection Equation					
Return	10.0				10.4099
					(0.0489)
Time	0.1				0.1072
					(0.0011)
Time-Squared	0.0				-0.0000
					(0.0000)
Constant	-1.0				-1.0247
					(0.0052)

Table B.IV: Convergence Tests: This table shows the Geweke (1992) and Gelman-Rubin (1992) tests for convergence, computed for our dataset of entrepreneurial firms in the paper, using a one-factor market model in monthly log returns, with selection correction (model 1 in table II). In the valuation equation, δ and β are the monthly intercept and the slope on the market log return (in excess of the risk-free rate). σ is the estimated monthly standard deviation of the error term. In the selection equation, γ_0 is the loading on the intercept, γ_1 is the loading on the log return since the previous financing event, and γ_2 and γ_3 are loadings on the time since the last financing event and its squared value. The Geweke (1992) Z-statistic is a difference in means test. After discarding the first 1,000 cycles, we calculate the mean and standard deviation of the first 10% (Mu1 and Sigma1) and the last 50% (Mu2 and Sigma2) of the next 5,000 cycles. We use Bartlett spectral density estimates of Sigma1 and Sigma2, to account for autocorrelation. We also report the p-values of the Z-statistic. The Gelman-Rubin (1992) R-statistic is based on 10 chains with 1,000 burn-in and 1,000 estimation cycles. Each chain has different starting values. We draw starting values of δ and β from a $N(0, 0.08^2)$ and $N(3, 1.5^2)$ distribution, respectively. The starting value for σ is drawn uniformly between 0 and 0.5. Starting values for γ are drawn from a $N(0, 0.5^2)$ distribution. Values of R-stat above 1.1 or 1.2 are usually considered non-stationary. For both convergence tests we use the priors described in appendix A.

	Geweke (1992)						Gelman-Rubin (1992)
	Mean 1	Std. dev. 1	Mean 2	Std. Dev. 2	Z-stat	p-value	R-stat
Valuation Equation							
Intercept	-0.0563	0.0020	-0.0563	0.0115	0.2028	0.8393	1.0274
RMRF	2.7675	0.2888	2.7709	0.4475	-0.1706	0.8646	1.0081
Sigma	0.4097	0.0108	0.4115	0.0656	-0.9886	0.3229	1.0426
Selection Equation							
Return	0.3334	0.0245	0.3316	0.0366	1.3207	0.1866	1.0655
Time	0.3725	0.1690	0.3698	0.0838	0.3506	0.7259	1.0361
Time-Squared	-0.03714	0.0244	-0.0365	0.0076	-0.6259	0.5314	1.0215
Constant	-1.9318	0.1218	-1.9302	0.0491	-0.2865	0.7745	1.0184

Table B.V: Robustness to Non-Normality of Error Term: This table reports MCMC estimates of the one-factor market model in monthly log returns, with selection correction. All reported estimates are mean and standard deviations (in parentheses) of the simulated posterior distributions. In the valuation equation,

$$v(t) = v(t-1) + r + \nu + \beta\left(r_m(t) - r\right) + \varepsilon(t),$$

δ and β are the monthly intercept and the slope on the market log return (in excess of the risk-free rate). The error term in the observation equation, ε, is a mixture of K Normals, with probability density $f_\varepsilon = \sum_{i=1}^{K} p_i N(\mu_i, \sigma_i^2)$. The priors on the mixture parameters $\mu_i \mid \sigma_i^2 \sim N(0, 100 \cdot \sigma_i^2)$ and $\sigma_i^2 \sim IG\left(2.1, 1/600\right)$ are the same as for the single Normal distributed error term in the paper. The parameter $\delta = \nu + \sum_{i=1}^{K} p_i \mu_i$ incorporates the mean of the mixture distribution. We report the moments of the centered error term $\varepsilon - \sum_{i=1}^{K} p_i \mu_i$, where kurtosis is in excess of the Normal distribution kurtosis. In the selection equation, γ_0 is the loading on the intercept, γ_1 is the loading on the log return since the previous financing event, and γ_2 and γ_3 are loadings on the time since the last financing event and its squared value. The simulations use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. ***, **, and * denote whether zero is contained in the 1%, 5%, and 10% credible intervals, respectively.

	K=1	K=2	K=3
Valuation Equation			
Intercept	-0.0563 ***	-0.0509 ***	-0.0508 ***
	(0.0016)	(0.0016)	(0.0017)
RMRF	2.7510 ***	2.7029 ***	2.6367 ***
	(0.1127)	(0.1134)	(0.1335)
Selection Equation			
Return	0.3321 ***	0.3266 ***	0.3284 ***
	(0.0079)	(0.0090)	(0.0089)
Time	0.3666 ***	0.3499 ***	0.3476 ***
	(0.0202)	(0.0211)	(0.0216)
Time-Squared	-0.0361 ***	-0.0352 ***	-0.0345 ***
	(0.0028)	(0.0029)	(0.0028)
Constant	-1.9290 ***	-1.9228 ***	-1.9203 ***
	(0.0170)	(0.0164)	(0.0170)
Error Term			
Mean	0.0000	0.0000	0.0000
Std. Dev.	0.4109	0.4064	0.4073
Skewness	0.0000	0.0011	0.0024
Kurtosis	0.0000	0.0060	0.0134