Risk and Return Characteristics of Venture Capital-Backed Entrepreneurial Companies

Arthur Korteweg, Stanford University, Graduate School of Business

Morten Sorensen, Columbia Business School, NBER, and SIFR

January 2011

Acknowledgements: We are particularly grateful for our discussions with John Cochrane that helped shape our understanding of the underlying problems, and we thank Anat Admati, Tom Davidoff, John Heaton, Josh Lerner, John Quigley, Caroline Sasseville, Annette Vissing-Jørgensen, two anonymous referees, and seminar participants at the University of Amsterdam, Columbia University, NYU Stern School of Business, University of Chicago, University of Illinois–Urbana Champaign, Copenhagen University, London Business School, the Stanford-Berkeley joint seminar, Tilburg University, the University of Iowa, Virginia Tech, Society of Quantitative Analysts, Q-group, UBC Sauder School of Business, and the 2009 American Finance Association Meetings in San Francisco for helpful comments. Susan Woodward of Sand Hill Econometrics provided generous access to data, and the Center for Research in Security Prices (CRSP) and the Kauffman Foundation provided financial support.

Contact Information: Arthur Korteweg, Stanford Graduate School of Business, 518 Memorial Way, Stanford, CA 94305. (650) 498-6993, korteweg_arthur@gsb.stanford.edu. Morten Sorensen, Columbia Business School, 3022 Broadway, New York, NY 10027, (212) 851 2446, ms3814@columbia.edu

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. The methodology is generally applicable to estimating risk and return in illiquid markets with endogenous trading.

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 non-synchronously 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). Moreover, when the timing of observed returns is endogenous, a sample selection problem arises. Here we address the latter problem, which we term the *dynamic selection problem*, as formally defined below.

The dynamic selection problem is pervasive, arising in areas as diverse as hedge funds, real estate, and venture capital or private equity investments. For hedge funds, numerous papers have studied the selection issues arising from the voluntary reporting of hedge fund performance data (e.g., Baquero, ter Horst, and Verbeek, 2005; ter Horst and Verbeek, 2007; and Jagannathan, Malakhov, and Novikov, 2009). Hedge funds with worse performance are more reluctant to report returns and less likely to survive, and the resulting self-selection and survivorship problems are manifestations of the dynamic selection problem. In real estate, transaction prices are only observed for traded properties, and the dynamic selection problem arises if, say, sellers with higher reservation prices are less likely to sell or properties that have depreciated more are more likely to trade, for example, due to foreclosures (Gatzlaff and Haurin, 1997; Fisher, Gatzlaff, Geltner, and Haurin, 2003; Hwang and Quigley, 2003; and Goetzmann and Peng, 2006).

Our study focuses on venture capital (VC) investments in entrepreneurial companies. Here the dynamic selection problem arises because valuations of portfolio companies are only observed when the companies receive funding or have exit events (IPOs or acquisitions). These events are more frequent for well-performing companies and these companies are more likely to subsequently survive. We find that the dynamic selection problem is important and that controlling for selection substantially decreases the estimated returns and increases the measures of the riskiness of entrepreneurial investments. In our baseline specifications, the estimated alpha decreases by about 40% and the market beta increases by about 20% relative to conventional GLS estimates that ignore the selection problem.

Our empirical approach extends existing empirical models of the risk and return of VC investments (Cochrane, 2005; and Hwang, Quigley, and Woodward, 2005). We extend a standard dynamic asset-pricing model by adding a selection process to correct for the endogenous selection of the observed returns. Our model explicitly specifies the entire unobserved valuation and return path between the observed valuations, as well as the probability of observing a valuation at each point in time. Formally, we combine a Type-2 Tobit model (Heckman, 1979; and Amemiya, 1985) with a dynamic filtering and smoothing problem (Kalman, 1960; and Anderson and Moore, 1979). We present a Markov Chain Monte Carlo estimator using Gibbs sampling (Gelfand and Smith, 1990; and Robert and Casella, 2004), which 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. Although we primarily report point estimates, this algorithm generates the posterior distribution of all the parameters and latent variables in the model, in particular it produces the estimated paths of all the unobserved valuations.

Our approach produces substantially different results than previous studies, primarily in our estimates of systematic risk. Cochrane's (2005) 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 of 2.8. 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 VC funds), and Gompers and Lerner (1997) report betas from 1.08 to 1.4 (using a sample of 96 VC 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. Overall, these results suggest that entrepreneurial investments are more risky than previously found.

The differences in the parameter estimates are likely due to the richer specifications that our approach accommodates. 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, and without such variation the estimated parameters may be sensitive to functional and distributional assumptions. We are able to include the time since the previous financing round in the selection equation as a reasonable source of exogenous variation, and we confirm in the

Appendix that our results are insensitive to distributional assumptions. Moreover, we can include Fama and French (1995) factors, a VC-specific factor, and period- and stage-specific parameters to capture company-level heterogeneity. In our three-factor specification, the loading on the size factor (SMB) turns positive, and the loading on the value factor (HML) increases substantially when controlling for selection. Perhaps not surprisingly, the risk profile of privately-held entrepreneurial companies resembles that of small, high-growth, and high-risk public companies. The exposure to market risk increases with the stage of investment, consistent with the notion that market conditions are important for VC's valuations. We also find evidence of a strong VC-specific factor defined in terms of the aggregate volume of VC investments. This is consistent with previous studies that suggest that capital inflows into VC funds help drive up valuations of entrepreneurial companies and in turn reduce VC investors' returns. We are also able to look at risk-adjusted returns by sub-period, and find moderate alphas during the period before 1995, whereas the 1995-2001 period is characterized by high alphas. Post-2001 the alpha appears to have turned negative.

Our approach allows us to test the robustness of our results in several new ways, detailed in the Appendix. First, we use flexible distributions of the error terms, and report estimates using mixtures of up to four normal distributions, finding results that are largely insensitive to this distribution and consistent with the more restrictive assumption in our main specification. Using simulated data, we show that our procedure is robust to misspecifications of the error distribution. Moreover, we estimate specifications with company-specific parameter heterogeneity using hierarchical priors, corresponding to a random coefficients model. Again, our results are largely insensitive to this relaxation,

supporting the more restrictive assumption of homogenous parameters used in the main specifications. A final advantage of our approach is that it delivers accurate finite-sample inference, even for non-linear functions of non-Gaussian parameters, such as the alpha in our model. 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.

A related literature estimates the risk and return of private equity and VC investments using the cash flows distributed to the limited partners (Gompers and Lerner, 1997; Jones and Rhodes-Kropf, 2003; Ljungqvist and Richardson, 2003; Kaplan and Schoar, 2005; Phalippou and Gottschalg, 2009; 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.

The paper is organized as follows. Section 1 provides a formal definition of the dynamic selection problem. Section 2 describes the econometric model and estimation algorithm, and Section 3 describes the data. Section 4 discusses the empirical results. In Section 5 we discuss the interpretation of the intercepts in the factor models, and Section 6 concludes. The Appendix describes the robustness and convergence properties of this

algorithm. An online appendix contains a more detailed description of the estimation procedure along with computer code to implement this procedure (http://XXXXX).

1. The Dynamic Selection Problem

To fix ideas and notation, the *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 problem arises when $\gamma_v \neq 0$, because $E[\varepsilon(t) \mid data] \neq 0$, conditioning on all observed data. Intuitively, the problem arises whenever the probability of observing a valuation is related to the valuation itself. In our application, entrepreneurial companies with higher valuations are more likely to be refinanced, so companies with higher realizations of ε are overrepresented in the data relative to the population, and $E[\varepsilon(t) \mid data] > 0$.

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 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 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:

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

This conditional mean is a function of the trading history and observables over the entire period between the observed valuations. Unlike the standard selection model,

observations with unobserved valuations are informative about the valuation process, and the conditional means for the observed valuations depend on the observables over the periods where the valuations were unobserved. 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.

2. Econometric Model and Estimation Procedure

To motivate our 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.

2.1 Valuation Process

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

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

The value of the market portfolio 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 excess return of the valuation process 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:

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

where $\varepsilon(t,t')$ is follows the $N(0,(t'-t)\sigma^2)$ distribution.[1] Defining $v(t) = \ln[V(t)]$, and starting from $t = t'-1$, we arrive at the one-period transition equation for the valuation 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 $r_m(t) = \ln[M(t)/M(t-1)]$. This is equation (1) with $X(t) = [1 \quad r_m(t) - r]'$ and $\theta = [r+\delta \quad \beta]'$.

2.2 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 $\eta(t) \sim N(0,1)$, and the parameters of interest are δ, β, σ^2, and $\gamma = (\gamma_0, \gamma_v)$.

2.3 Overview of Estimation Procedure

We use a Bayesian Gibbs sampling procedure (see Geman and Geman, 1984; Tanner and Wong, 1987; Gelfand and Smith, 1990; and Johannes and Polson, 2010), which allows us to divide our model into three blocks. The first one contains the valuation variables. Most valuations are unobserved in the data, and this block estimates

the unobserved valuations, which are treated as parameters within the model. The second block 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 is similar to Bayesian estimation of a probit model (Albert and Chib, 1993). For the parameters in the valuation and selection equations in the third block, we use two standard Bayesian linear regressions, given by equations (11) and (13).

Drawing the valuation variables in the first block is the most complex part of the procedure. This part traces 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. The starting values and prior distributions are provided in the Appendix. A more detailed description of the procedure along with code to implement our procedure is available in an Online Appendix.

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, the selection variables $w(t)$ can be regarded as noisy "observations" of $v(t)$, and the first observation equation is the selection equation (13). 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 would be possible to incorporate observation error here without losing the linear filtering properties.

We use diffuse priors and several different starting values for the parameters, as detailed in the Online Appendix. 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 the Appendix, including relaxing the assumptions of Gaussian error terms and that alpha and beta are constant across companies.

3. Data Description

Monthly market returns and returns to Fama-French portfolios (Fama and French, 1995) 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 his website.

3.1 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 VC investments in the United States 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 previously reported data problems have been resolved.[2]

3.2 Calculating Returns

VCs distinguish between pre- and post-money valuations. 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, the gross return earned by an investor over two subsequent rounds from time t to t' is:

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

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.[3]

3.3 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, the data are more likely to include valuations for companies with IPOs or acquisitions, since valuations of IPOs and acquisitions are publicly available. Consequently, these companies are overrepresented in the sample of companies for which we have return information. To adjust for this oversampling, we use a randomly drawn subsample that matches the IPO and acquisition rates of the full dataset.[4] The number of companies and their exits, in the full dataset and in the sample used for estimation, are listed in Table 1.

**** TABLE 1: 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.[5]

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%.

4. Risk Factors for Entrepreneurial Companies

Table 2 presents four different specifications of the selection equation. The valuations appear highly exposed to the market factor with a beta (RMRF) around 2.8. The (monthly) intercept is about -5.7%, and the (monthly) standard deviation of the idiosyncratic returns (*Sigma*) is 41%. Note that the intercept is not an abnormal return, but it is possible to compute the implied posterior distribution of alpha using $\alpha = \delta + \frac{1}{2}\sigma^2 - \frac{1}{2}\beta(1-\beta)\sigma_m^2$. We return to this calculation below. The 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 smaller than the betas for the companies in our sample.

**** TABLE 2: ONE-FACTOR MODEL ****

In the first specification in Table 2, 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 five 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 be refinanced 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. It is well known that a valuation process given by $V(t) = E[V_T \mid F(t)]$ is a martingale, where V_T is the final payoff and $F(t)$ is a filtration, and consequently the error terms in this process are independent over time. In particular, they are independent of *Time*, which is the exclusion restriction. Moreover, *Time* is directly related to the probability of observing a financing round. VC financing involves sufficient capital to sustain the company for a substantial period, and a company is unlikely to be refinanced repeatedly in short succession. After the company exhausts its capital, typically after one or two years, the probability of refinancing increases. If too much time elapses, however, the company is likely struggling, and the probability of being refinanced declines again. In this case, *Time* is a valid source of exogenous variation for semi-parametric identification of the model.

The critical assumption is that the investor's valuation process can be written as $V(t) = E[V_T \mid F(t)]$. This states that today's valuation is the expected future payoff, rationally discounted and risk-adjusted taking into account all current information. The assumption is standard, and the main requirement is that investors rationally anticipate future contingencies. The condition remains valid even with illiquid and opaque markets for entrepreneurial companies, with future negotiations between entrepreneurs and investors, and with asymmetric information and high uncertainty about future performance, as long as these contingencies are rationally anticipated by the VCs ex-ante. We cannot test whether VCs are fully rational. For public markets, there is substantial support for the efficient market hypothesis, suggesting that valuations incorporate current information about future contingencies. We have no reason to expect that VCs should be less sophisticated than public market investors.

In the second specification of the selection process in Table 2, 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 2, 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 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 appear to be 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 VC-specific component below.

4.1 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 this bias. Table 3 presents estimates of these models. For the standard OLS and GLS estimators, we calculate the log excess returns and regress them on the corresponding log excess market returns. In particular, 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 intercept in equation (10) and is called *Intercept* here as well although, strictly speaking, equation (16) has no intercept. When observed valuations are more distant in time they have more volatile errors, however, 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 (10) and is called *Intercept* here as well.

**** TABLE 3: OLS, GLS, AND MCMC ****

Comparing the OLS and GLS estimates, we see 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 3 also ignore selection (by setting $\gamma_v = 0$, see details in Online Appendix). Comparing these MCMC specifications to the specifications in Table 2 with selection corrections, we find that the intercept increases from -5.7% to -1.6% to per month for the procedure that does not correct for selection.[6] The change in beta (*RMRF*) is smaller, decreasing from 2.75 to

2.66 without selection correction, and the estimated volatility declines from a monthly standard deviation of 41% to 36%. 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.[7]

4.2 Three-Factor Model

Table 4 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*). We still 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 factor ranging from 1.22 to 1.47. Fama and French (1995) 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, indicating that for VC-backed private companies, growth options represent a much 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 explain variations in the returns to privately-held entrepreneurial companies.

**** TABLE 4: THREE-FACTOR MODEL WITH SELECTION ****

4.3 Comparing Companies Across Stages and Periods

Table 5 presents estimates with separate coefficients for investments in companies at different stages. We refer to four stages of development: "seed," "early," "late," and "mezzanine," as defined by Sahlman (1990). 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. This is consistent with the definition of seed investments, which are primarily investments to develop young ideas or prototypes where the risk is mainly idiosyncratic technological risk. 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.

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 5: ESTIMATES BY COMPANY STAGE ****

In Table 6 and Figure 3, 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 VC 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 1990s have abated in recent years.

**** TABLE 6: ESTIMATES BY INVESTMENT PERIOD****

4.4 Estimates with VC-Specific 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 VC funds lead to higher valuations and subsequent poorer performance. This effect may introduce a risk factor that is specific to VC investments, and our factor is an attempt to provide a measure of the potential magnitude of this risk. In Table 7, we see that the valuations load strongly on the VC factor, suggesting that there may be 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 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 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 7: ESTIMATES WITH VC FACTOR ****

5. 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 correction for multi-factor specifications is in footnote 1. One advantage of the Bayesian approach is that it is

possible to compute the posterior distribution of alpha even though it 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 8. We first report GLS and MCMC estimates without correcting for selection. Without correcting for selection, we see high monthly alphas between 5.2% and 7.9%.

Correcting for selection, the alphas drop substantially, as indicated both in Table 8 and Figure 2. Figure 2 compares the posterior distributions of the alphas found using the MCMC estimates without selection correction from Table 3 and the estimates with correction from the first specification in Table 2. Correcting for selection, the market model specifications in Table 2 and the Fama-French specifications in Table 4 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 8 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%. This increased dramatically in the late 1990s, during the dot-com boom, to a monthly alpha around 5.8%. The 2001-2005 period 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 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 VC

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 specific 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 the 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 8 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 8: ESTIMATES OF ALPHA ****

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

6. 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 and specifications of 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 to address these problems, and it may be possible to apply our methodology, with some modifications, to those areas as well.

Appendix: Robustness and Convergence

Below we provide details about the prior distributions and starting values of the sampling procedure and confirm the robustness of our procedure by using simulated data, testing for convergence, relaxing the normality assumption, and allowing for firm-level heterogeneity in the parameters.

A.1 Prior Distributions and Starting Values

We use diffuse priors for the parameters. We set the prior means of δ, β, and (γ_0, γ_v) 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 inverse gamma 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 $N(0,10^2)$.[8] We start the Markov chain with δ, β, and (γ_0, γ_v) 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 $w(t)$.

We implement this algorithm in C++, using the GNU Scientific Library (GSL). On a 2.66 GHz Pentium 4 quad-core processor, it takes about 30 minutes to simulate 6,000 draws of the Markov Chain (using only a single core).

A.2 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. 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 9 to 11. 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 in parentheses.[9] Overall, our algorithm seems to recover the underlying parameters well, even with 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.

A.3 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 2. Convergence tests for the other specifications and the simulated data produce similar results.

Figure 4 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 12 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 12, 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.

A.4 Relaxing 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. 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:

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

which 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, is 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 $\mu_i \mid \sigma_i^2 \sim N(0, 100 \times \sigma_i^2)$. Figure 6 presents parameter plots and the estimated parameters are in Table 13. In the top plot of Figure 6, 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 the second plot. Similar patterns are observed for the variances and probabilities, as seen in the

bottom two plots. In Table 13, 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.

A.5 Company-Specific Coefficients Using Hierarchical Priors

As a final robustness check, we investigate whether our results are sensitive to the assumption that intercepts and betas are constant across companies. We estimate company-specific δ_i and β_i using a hierarchical prior approach, similar to a random coefficients specification [for additional details, see Rossi, Allenby, and McCulloch (2005)]. We specify that (δ_i, β_i) is distributed *i.i.d* across companies with:

$$(\delta_i, \beta_i) \sim N\left((\delta_0, \beta_0), S\right), \tag{19}$$

where δ_0, β_0, and S are parameters to be estimated. For these parameters, we use the priors:

$$S \sim IW(h, H) \tag{20}$$

$$(\delta_0, \beta_0) \mid S \sim N\left((0,0), 100 \times S\right) \tag{21}$$

where $IW(h, H)$ denotes the conjugate Inverse Wishart distribution. Following Rossi, Allenby, and McCulloch (2005), we set $h = 5$ and

$$H = (h-3)\begin{bmatrix} 0.1^2 & 0 \\ 0 & 1^2 \end{bmatrix}, \tag{22}$$

implying a mean of S of $\begin{bmatrix} 0.1^2 & 0 \\ 0 & 1^2 \end{bmatrix}$. For each company the prior standard deviations of δ_i and β_i are 0.1 (10% per month) and 1, respectively. For the hyperparameters, the standard deviations of δ_0 and β_0 are ten times as large, i.e., 1 and 10. The priors on the idiosyncratic volatility and the parameters in the selection equation are left unchanged.

We extend the Gibbs procedure to sample δ_i and β_i from company-by-company regressions with "priors" (δ_0, β_0) and S. Given these draws, we draw (δ_0, β_0) and S from the posterior distribution of the multivariate regression of δ_i and β_i on a constant with priors given by equations (20) and (21). Given the increased dimensionality of this model, particularly the firm-specific deltas and betas, we use 20,000 iterations for estimation, discarding the initial 10,000 ones for burn-in.

We estimate the hierarchical version of specification 1 of Table 2. The top plots of Figure 7 show the posterior distribution of δ_0 and β_0. For δ_0, the posterior distribution has mean -0.0572 and standard deviation 0.0017. For β_0, the mean is 2.8272 and the standard deviation is 0.1335. The idiosyncratic volatility has mean 0.4032 and standard deviation 0.0065. The average alpha across firms is 0.0399. Note that the standard deviations of δ_0 and β_0 are fairly small, and their distributions are very similar to the means and standard deviations reported in specification 1 in Table 2. Overall, this

suggests that parameter heterogeneity across companies is not a substantial concern for our estimates.

References

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

Amemiya, T. 1985. *Advanced Econometrics*. Harvard University Press, Cambridge, MA.

Anderson, B. and J. Moore. 1979. *Optimal Filtering*. Prentice Hall, New York, NY.

Andrews, D. and M. Schafgans. 1998. Semiparametric Estimation of the Intercept of a Sample Selection Model. *Review of Economics Studies* 65:497-517.

Baquero, G., J. ter Horst, and M. Verbeek. 2005. Survival, Look-Ahead Bias, and Persistence in Hedge Fund Performance. *Journal of Financial and Quatitative Analysis* 40:493-517.

Campbell, J., A. Lo, and A. MacKinlay. 1997. *The Econometrics of Financial Markets.* Princeton University Press, Princeton, NJ.

Carter, C., and R. Kohn. 1994. On Gibbs Sampling for State Space Models. *Biometrika* 81:541-553.

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

Davis, J., E. Fama, and K. French. 2000. Characteristics, Covariances, and Average Returns: 1929 to 1997. *Journal of Finance* 55:389-406.

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

Driessen, J., T-C. Lin, and L. Phalippou. 2007. A New Method to Estimate Risk and Return of Non-Traded Assets from Cash Flows: The Case of Private Equity Funds. Working paper, University of Amsterdam.

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

Fisher, J., D. Gatzlaff, D. Geltner, and D. Haurin. 2003. Controlling for the Impact of Variable Liquidity in Commercial Real Estate Price Indices. *Real Estate Economics* 31:269-303.

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

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

Gelfand, A., 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, S., and D. Geman. 1984. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721-741.

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

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

Gompers, P., and J. Lerner. 1997. Risk and Reward in Private Equity Investments: The Challenge of Performance Assessment. *Journal of Private Equity* 1:5-12.

Gompers, P., and J. Lerner. 1999. *The Venture Capital Cycle*. MIT Press, Cambridge, MA.

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

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

ter Horst, J., and M. Verbeek. 2007. Fund Liquidation, Self-Selection, and Look-Ahead Bias in the Hedge Fund Industry. *Review of Finance* 11:605-632.

Hwang, M., and J. 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, M., J. Quigley, and S. Woodward. 2005. An Index for Venture Capital, 1987-2003. *Contributions to Economic Analysis & Policy* 4:1-43.

Jagannathan, R., Alexey M., and D. Novikov. 2009. Do Hot Hands Exist among Hedge Fund Managers? An Empirical Evaluation. *Journal of Finance,* Forthcoming.

Johannes, M., and N. Polson. 2010. MCMC Methods for Financial Econometrics. In: Y. Ait-Sahalia and L. Hansen (eds.) *Handbook of Financial Econometrics,* Forthcoming.

Jones, C., and M. Rhodes-Kropf. 2003. The Price of Diversifiable Risk in Venture Capital and Private Equity. Working paper, Columbia University.

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

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

Kaplan, S., B. Sensoy, and P. Strömberg. 2002. How Well Do Venture Capital Databases Reflect Actual Investments? Working paper, University of Chicago.

Ljungqvist, A., and M. Richardson. 2003. The Cash Flow, Return and Risk Characteristics of Private Equity. Working paper, NYU Stern School of Business

Peng, L. 2001. Building a Venture Capital Index. Working paper, University of Boulder, Colorado.

Phalippou, L., and O. Gottschalg. 2009. The Performance of Private Equity Funds. *Review of Financial Studies* 22:1747-1776.

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

Robert, C., and G. Casella. 2004. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, NY.

Rossi, P., G. Allenby, and R. McCulloch. 2005. *Bayesian Statistics and Marketing*. John Wiley and Sons, Chichester, UK.

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

Scholes, M., and J. Williams. 1977. Estimating Betas from Nonsynchronous Data. *Journal of Financial Economics* 5:309-328.

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

[1] 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.

[2] 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 filters the data to account for outliers. In contrast, we only had to eliminate a single round in which the return was below -100%. In the Appendix we relax the normality assumption to allow for fat tails and skewed distributions but find no evidence of outliers in our data.

[3] 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.

[4] For each company in the sample that goes public (is acquired), it is included in the subsample with probability p_i/q_i, where p_i is the frequency of companies going public (being acquired) in the full dataset, and q_i is the frequency in the sample with returns. Using the full sample of observed valuations changes the intercept and beta in the outcome equation from -0.0563 and 2.7510 to -0.032 and 2.7886, and has negligible effects on the other parameters. Here, as below, we assume that the subsample of companies with valuation information is random conditional on the observed exit. If this were not the case, it would introduce an additional distinct sample selection problem.

[5] Our results are robust to this assumption. Extending the period to 120 months, the estimates in specification 1 in Table 2 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 less affected.

[6] Note that a formal comparison of these models is complicated by the fact that frequentist and Bayesian estimates are not directly comparable.

[7] 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.

[8] 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.

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

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

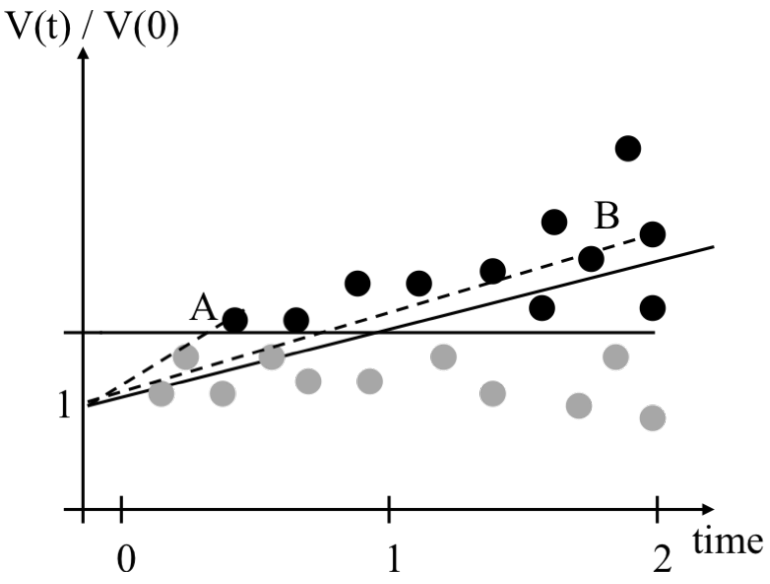

Figure 2: Illustration of selection bias on estimates of intercept, systematic risk, and idiosyncratic volatility.

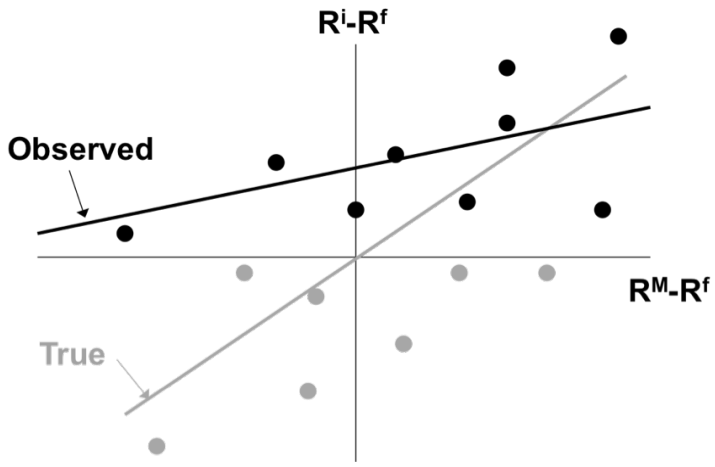

Figure 3: Posterior distribution of monthly excess return: This figure plots the posterior distribution of monthly risk-adjusted excess returns, α, based on the one-factor market model in log-returns. In the top plot, we estimate the model using an MCMC algorithm that uses the information in the selection equation to adjust for dynamic selection (specification 1 in Table 2) and an MCMC algorithm that ignores the information in the selection equation (specification 1 in Table 3). In the bottom plot, we plot the distribution of α by sub-period, as described in Table 6, using specification 1.

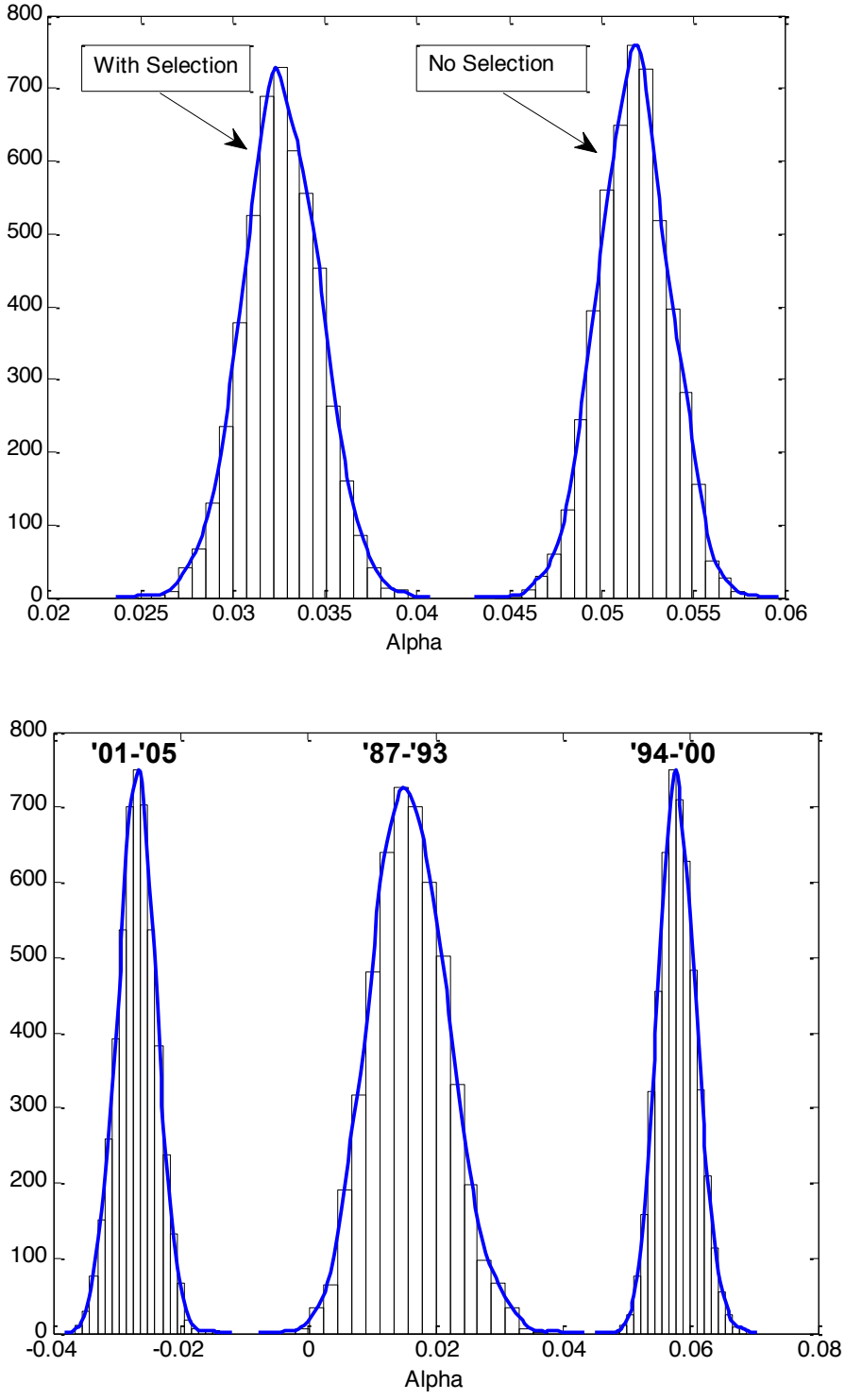

Figure 4: Trace plots: Plots of the parameter draws from the MCMC estimation of the one-factor market model in monthly log returns, with selection correction (model 1 in Table 2). In the valuation equation, δ and β are the monthly intercept and the slope on the market log return (in excess of the risk-free 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.

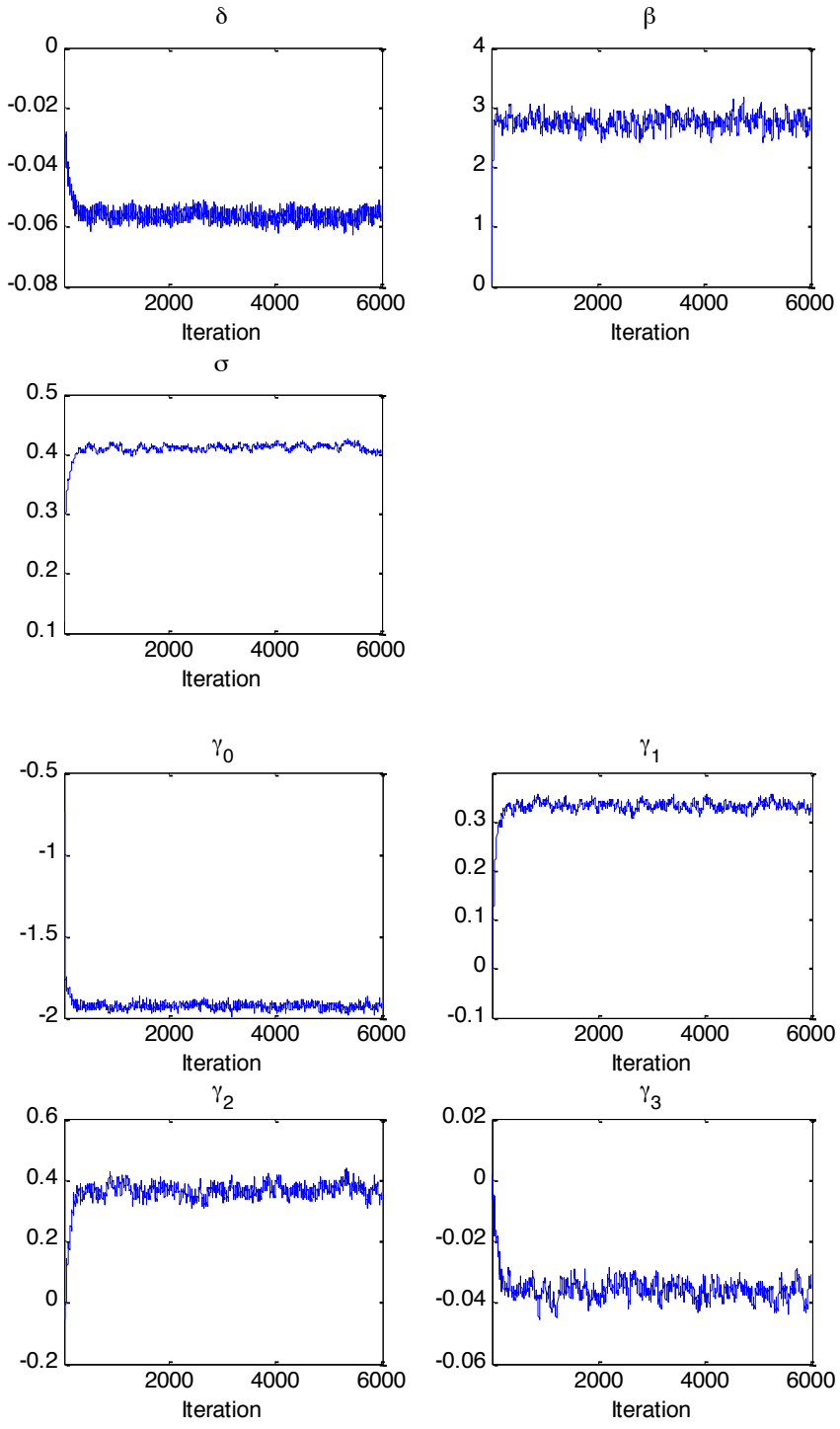

Figure 5: Auto-correlation functions: Plots of the autocorrelation functions of the MCMC draws of the one-factor market model in monthly log returns, with selection correction (model 1 in Table 2). The autocorrelations are calculated from 5,000 iterations of the MCMC algorithm, after discarding the first 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.

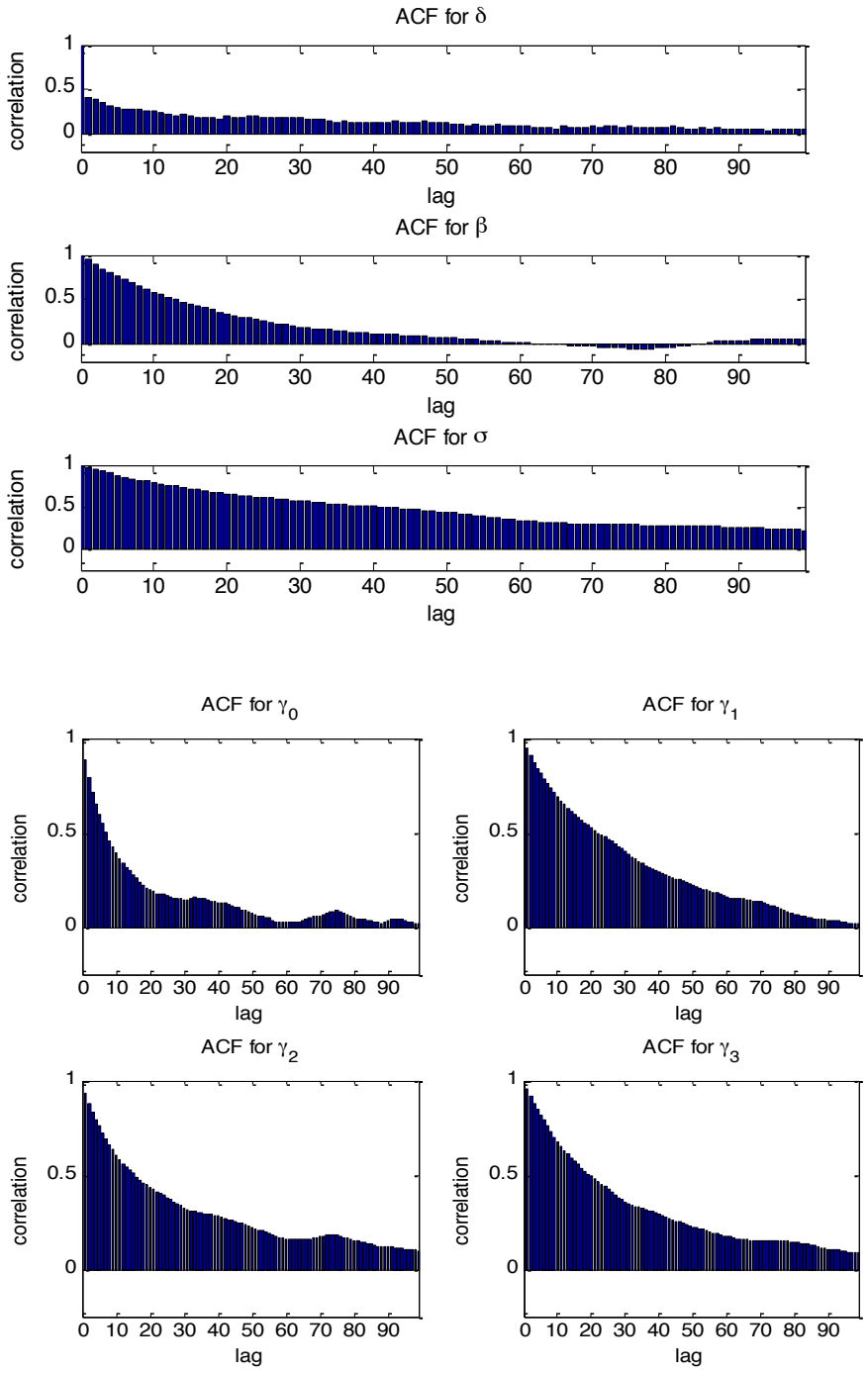

Figure 6: Trace plots of error term parameters for mixture of 2 normals: Plots of parameter draws of the MCMC estimation of the mixture model with 2 normals in the error term. We graph the individual error mixture distribution means (μ_1 and μ_2) along with the valuation equation intercept (ν) in the 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, σ_1 and σ_2 , are in the third plot, and the last plot shows the probabilities of the mixture distributions.

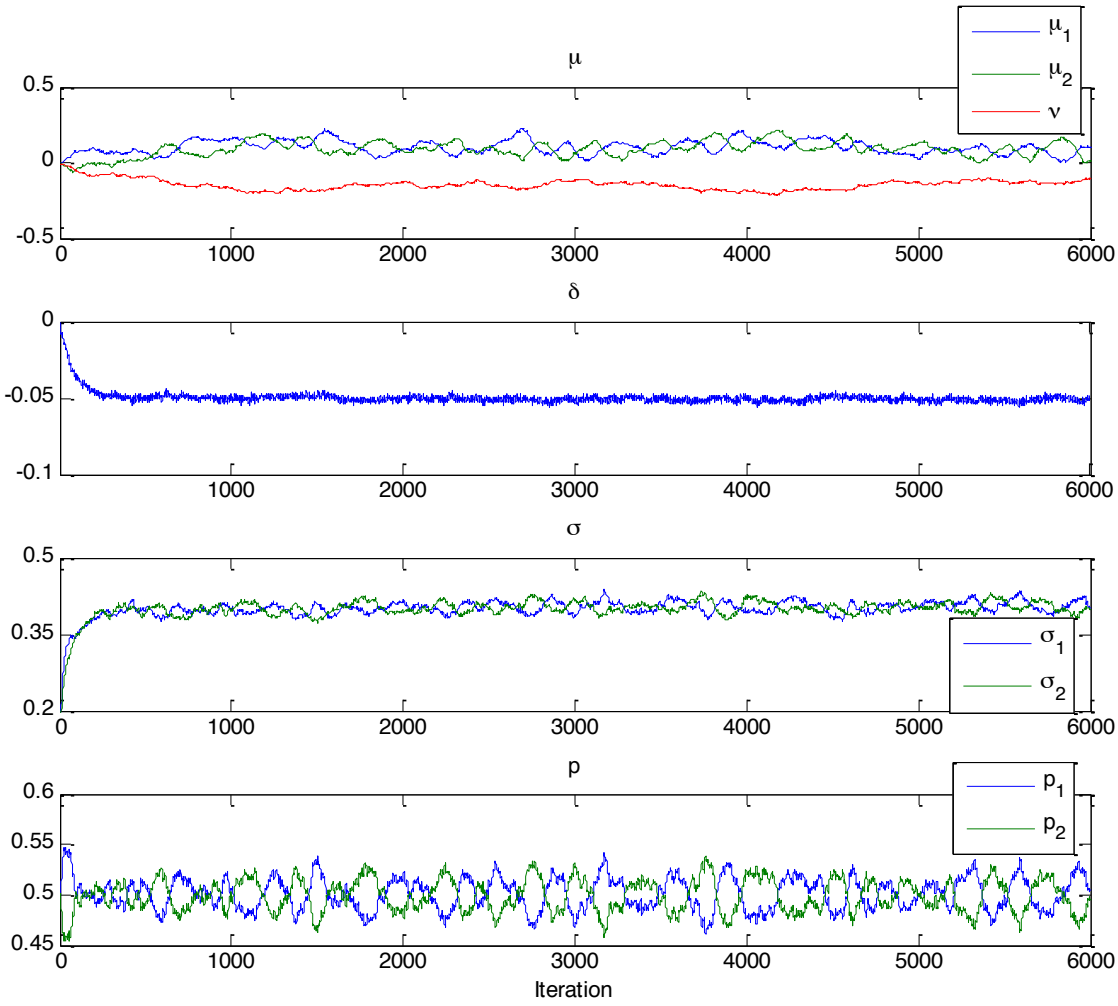

Figure 7: Cross-sectional distribution of risk and return in hierarchical model: Histograms of the distribution of firm-specific risk and return estimates from the MCMC estimation of the one-factor market model with hierarchical priors, as described in the Appendix. In the valuation equation, δ_i and β_i are firm-specific (monthly) intercept and the slope on the market log return (in excess of the risk-free rate), and α_i is the risk-adjusted excess return. The plots show the histograms of the posterior means of δ_i, β_i, and α_i across the 1,934 firms in the sample. The posterior means are calculated from 10,000 iterations of the MCMC sampler (after discarding the first 10,000 iterations).

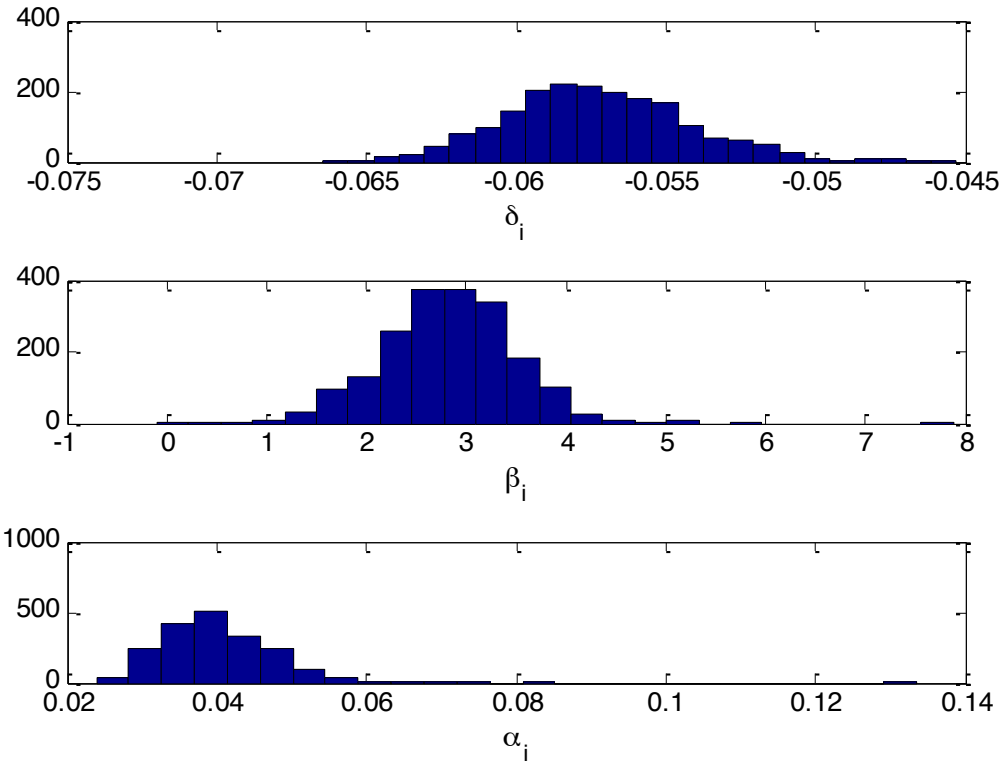

Table 1: Descriptive statistics: This table describes the number of rounds and companies, along with their exits, in the full dataset and the subsample used for estimation. The full dataset is obtained from Sand Hill Econometrics. The subsample used for estimation contains only companies with valuation information adjusted to match the IPO and acquisition rates in the full data, as described in the text.

	Data	Subsample
Rounds	61,356	5,501
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 2: 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. The estimates are means and standard deviations (in parentheses) of the simulated posterior distributions. In the valuation equation, *Intercept* is the monthly intercept in excess of the risk-free rate and *RMRF* is the slope on the market log return in excess of the risk-free rate. *Sigma* is the estimated 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). *Acquisitions, IPOs,* and *Rounds* contain the number of VC-backed acquisitions, IPOs, and total VC investment rounds in the month of the observation (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)
Sigma	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 3: OLS, GLS, and MCMC estimates: The table presents OLS, GLS, and MCMC estimates of the market model and the Fama-French three-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 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 the 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 standard deviation of the error term. The MCMC estimator use 5,000 iterations preceded by 1,000 discarded iterations for burn-in. For OLS and GLS estimates ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively. For Bayesian estimates, they denote 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 4: Estimates of Fama-French three-factor model with selection correction: 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. The estimates are means and standard deviations (in parentheses) of the simulated posterior distributions. In the valuation equation, *Intercept* is the monthly intercept in excess of the risk-free rate and *RMRF* is the slope on the market log return in excess of the risk-free rate. *SMB* is the small-minus-big portfolio, and *HML* the high-minus-low book-to-market portfolio. *Sigma* is the estimated 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). *Acquisitions, IPOs,* and *Rounds* contain the number of VC-backed acquisitions, IPOs, and total VC investment rounds in the month of the observation (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	-0.0538 ***	-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 5: Estimates by stage of development of entrepreneurial company: 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. The estimates are means and standard deviations (in parentheses) of the simulated posterior distributions. The specifications contain separate coefficients for companies at the seed, early, late, and mezzanine stages, as defined in Table 2 in Sahlman (1990). Our late stage corresponds to the second, third, and fourth stage according to Sahlman's definition. In the valuation equation, *Intercept* is the monthly intercept in excess of the risk-free rate and *RMRF* is the slope on the market log return in excess of the risk-free rate. *SMB* is the small-minus-big portfolio, and *HML* the high-minus-low book-to-market portfolio. *Sigma* is the estimated 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). *Acquisitions, IPOs,* and *Rounds* contain the number of VC-backed acquisitions, IPOs, and total VC investment rounds in the month of the observation (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 6: Estimates by investment period: 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. The estimates are means and standard deviations (in parentheses) of the simulated posterior distributions. The specifications contain separate coefficients for investments during the periods 1987-1993, 1994-2000, and 2001-2005. In the valuation equation, *Intercept* is the monthly intercept in excess of the risk-free rate and *RMRF* is the slope on the market log return in excess of the risk-free rate. *Sigma* is the estimated 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). 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 7: Estimates with VC factor: 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. The estimates are means and standard deviations (in parentheses) of the simulated posterior distributions. In the valuation equation, *Intercept* is the monthly intercept in excess of the risk-free rate and *RMRF* is the slope on the market log return 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 the total dollar volume of VC investments in the month of the observation. *Sigma* is the estimated 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). *Acquisitions, IPOs,* and *Rounds* contain the number of VC backed acquisitions, IPOs, and total VC investment rounds in the month of the observation (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 8: 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). See text for construction of these estimates.

	mean	std.dev.	1	5	50	95	99
Table 3: No selection							
Model 1							
GLS	0.0681						
MCMC	0.0517	(0.0019)	0.0472	0.0486	0.0517	0.0549	0.0562
Model 2							
GLS	0.0794						
MCMC	0.0560	(0.0022)	0.0513	0.0525	0.0559	0.0598	0.0613
Table 2: 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 4: 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 5: 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 6: 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 9: 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 $\varepsilon(t)$: $N(0, \sigma^2)$ and $\eta(t)$: $N(0,1)$ 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 the Appendix. Both MCMC methods use the same priors as specified the in the Appendix. 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 10: Estimates using simulated data with *t*-distributed errors: Simulations of the model described in Table 9, but with $\varepsilon(t)$: $0.0775 \cdot t_5$, 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 9 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 11: Estimates using simulated data with log-normal errors: Simulations of the model described in Table 9, but with $\varepsilon(t):\ LN(0,0.0993^2)-\exp(0.0993^2/2)$. 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 9 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 12: 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 2). 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 the Appendix.

	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 13: 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, 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* normal distribution, with probability density $f_\varepsilon = \sum_{i=1}^{K} p_i N(\mu_i, \sigma_i^2)$. The priors on the mixture parameters $\mu_i \mid \sigma_i^2 : N(0, 100 \cdot \sigma_i^2)$ and $\sigma_i^2 : IG(2.1, 1/600)$ 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