Inference and Prediction of Stock Returns using Multilevel Models

Brice Green[*]
Samuel Thomas[†]

2019-08-31

Abstract

Multilevel models are a generalized form of traditional linear regression models and have several benefits relative to traditional OLS regression including the regularization of parameter estimates, the ability to incorporate prior information, better out-of-sample forecasts, desirable inferential properties, and the ability to directly model the time-series/cross-sectional nature of financial security returns. We demonstrate that multilevel models generalize well-known asset pricing regression techniques like Fama-Macbeth and Fama-French regressions. They also have stronger explanatory power than traditional regression techniques, with a substantially lower out of sample mean squared error and a 5% to 10% higher out of sample R^2 vs. the comparable model fit with OLS.

Introduction

The search for factors that drive stock returns has produced close to 200 published variables that claim to be associated with higher returns. However, in the period following publication, many of the puported relationships either diminish in magnitude or disappear altogether (McLean and Pontiff 2016). Increasingly, financial economists are concerned that the high noise-to-signal ratio in security returns makes false discoveries altogether too easy. The procedures typically used to forecast risk and identify risk premia are subject to large numbers of researcher degrees of freedom (Gelman and Loken 2013), which have led to "statistically significant" findings that are a product of mis-aggregated noise (Harvey 2017). Putting stocks into portfolios and then performing regressions or T tests can be directly harmful to the power of the test statistics if the distribution of the relative ranks of the outcome variable is at all related to the sorting and splitting procedure for forming the portfolios (Lo and MacKinlay 1990).

Traditional approaches to forecasting financial risk and measuring risk premia have trouble coping with the vast amount of uncertainty in financial markets and large, idiosyncratic events. Market β, for example, is typically measured by regressing a security against a market portfolio over a window, but measurements long-run risk exposures can be swamped by temporary price movements that lead to spurious covariance estimates between the market and a given security. For example, news about fraudulent activity or fears of a changing regulatory environment might lead to temporary mispricing relative to its systematic risk, but the price could revert based on the security's longer-term fundamentals if the news proves to be temporary.

This paper addresses these concerns and provides a mechanism for forecasting both risk and prices, with the goal of mitigating false discoveries and improving out of sample forecasts. It does so by describing and then applying Bayesian multilevel models, a generalized class of regression models that aggregate relationships across heterogenous groups while allowing each group to have a varying relationship with the parameters. These models inherently regularize estimates towards a central value to an extent that depends on the heterogeneity of the underlying groups. The models generate two levels of parameter values: individual estimates at the group level that are useful for forecasting, and aggregate, "unconditional" parameter estimates that are useful for inference.

[*]Capital Group, bccg@capgroup.com
[†]Capital Group, sljt@capgroup.com
Many thanks to Brett Hammond and Steve Fox for useful comments.

Multilevel models are not yet common practice in financial economics, but they are regularly employed in educational research (Rubin 1981), political science (Ghitza and Gelman 2013), and meta-analysis (Meager 2019). The regularizing aspects of the "partial pooling" inherent in the structure of these models (averaging parameter estimates between in-group and across-subject loadings) help to mitigate the impacts of multiple comparisons by pulling an estimate towards its central tendency to the extent warranted by the between-group variance (Gelman, Hill, and Yajima 2012). They are also related conceptually to well-known econometric approaches like "fixed-effects" and "random-effects" models.

One reason that Bayesian multilevel models models have not been more widely adopted in finance is that until recently thye have been computationally intractable for models with large numbers parameters. Bayesian inference relies on the ability to marginalize out all of the parameters in the model, and thus requires the computation of an integral that usually has no closed-form solution. Recent advances in Markov Chain Monte Carlo (MCMC) techniques, such as Hamiltonian Monte Carlo (HMC) with No-U-Turn Sampling (NUTS), now allow researchers to fit the high-dimensional integrals necessary to estimate the model without having to specify various tuning parameters (M. D. Hoffman and Gelman 2014). Older MCMC methods like Metropolis-Hastings estimate expectations using a random walk, which takes an long time to converge to a typical set in high-dimensional spaces. By contrast, Hamiltonian Monte Carlo uses the gradient of the log posterior to guide proposals, increasing the efficiency of the sampler (Betancourt 2017). The flexibility of HMC and the automated tuning via NUTS means that academics and practitioners can easily target non-Gaussian generative distributions, fit models with large numbers of parameters, and choose priors for reasons other than the ability to easily fit the integral of interest.

From an inferential perspective, multilevel models are better able to control for unobserved differences between securities. Portfolio sorts assign stocks into a portfolio given a quantile of an underlying factor, but that assignment rule lacks the conditional independence assumptions needed to properly interpret the portfolio return as a treatment. A multilevel model uses each unit as its own control, aggregating the relationships across stocks as is warranted by the between-stocks variance. As a consequence, multilevel models provide a means to control for otherwise unobservable heterogeneity under relatively weak assumptions.

Because multilevel models pool information across stocks to come to estimates that aren't conditioned on group-specific information, they can make predictions about unobserved groups that are left out of the training data. This means that if we fit Fama-French style regressions for some of the 200 published factor models, we can compare those models through cross-validation (either k-fold or leave-one-out). This gives multilevel models a distinct advantage when it comes to model selection and weighting. If a given model has better predictive characteristics for securities left out of the training phase that gives a strong reason to prefer it. While in-sample R^2 consistently gets better as more parameters are introduced, cross-validation scores will get worse as nuisance parameters are added.

Bayesian models, generally, also allow for more complex types of hypothesis testing because they are fully generative. For example, a Bayesian multilevel model can *directly* test whether higher market β is associated with lower α. The correlation between the two parameters is estimated by the model, and its posterior distribution can be used for hypothesis testing at any arbitrary confidence interval.

The paper starts with an introduction to multilevel models and Bayesian inference, describes how these models can be applied to common financial models of stock returns, and demonstrates their performance relative to traditional techniques for forecasting risk and pricing factors. It then concludes with notes about new types of model comparison and hypothesis testing made possible by the modeling approach.

Multilevel Models

Multilevel models are a general class of regression models that extend traditional linear regression for data that has multiple levels of analysis with repeated measurements. For example, education researchers interested in the impact of both classroom and school-wide interventions on test scores need to account for the nested nature of the data. In these kinds of meta-analyses, we want to aggregate treatment effects across experiments, all the while knowing that each experiment has its own idiosyncracies. Being able to understand

how similar the different groups are is key to how and why multilevel models are valuable.

Specifically, given a set of groups $1, ..., j$ measured over time (be it stock returns or test scores). The time-series/cross-sectional data has two levels: the aggregated set of measurements might represent something about the class generally, while within each group we also have repeated measurements, so the groups could be modeled separately. If the groups are extremely different, we might not want to pool the estimates, but if they are highly similar we want the extra power that comes from pooling them together. Supposing that we had a series of measurements x for each group, classical models either pool all of these observations together or separate them entirely, and our estimates become $\alpha = E(x)$ for the "pooled" model and $\alpha_i = E(x_i)$ for the "no pooling" model.

A multilevel generalizes across both of these approaches by converging to a weighted average of the two parameters. The weight given to the across-group α versus group i's α_i is determined by the between-group variance σ_α^2, the within group variance σ_y^2, and the sample size for the given group n_i. The multilevel estimate can be approximated as

$$\alpha_i^{multilevel} = \frac{\frac{n_i}{\sigma_y^2}\overline{y_i} + \frac{1}{\sigma_\alpha^2}\overline{y_{all}}}{\frac{n_i}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}},$$

where $\overline{y_j}$ is the group level mean and $\overline{y_{all}}$ is the completely pooled estimate (Gelman and Hill 2006, 254).

In limiting cases, multilevel models reduce to well known estimators in the economics literature. For example, as σ_α^2 approaches 0, $\alpha_i^{multilevel}$ becomes the pooled average across all measurements, ignoring the differences between groups. In the case where σ_α^2 approaches ∞, the expression reduces to $\overline{y_i}$, which is the fixed-effects estimator in economics, a common approach to inference across groups that otherwise have unobserved differences. Fixed effect estimators require that heterogeneity in between groups can be captured properly by the group-level sample mean (that is to say that regression slopes are not varying), while multilevel models can allow for both group-level intercepts and slopes.

Bayesian Estimation

Multilevel models typically rely on Bayesian estimation because they are very difficult to fit without relying on markov chain monte carlo methods and prior information. There are frequentist statistical packages like the R package lme4 (Bates et al. 2007) which rely on restricted maximum likelihood and optimization techniques to arrive at parameter point estimates. Optimization-based maximum likelihood approaches also require strong assumptions about the distribution of the group-level effects in order to identify the model, and in high-dimensional settings with many groups and parameters, REML-based techniques often fail to converge.

Bayesian inference relies on Bayes' rule to arrive at a posterior distributions for the model parameters. Suppose we have some data and a model with k parameters $\Theta = (\theta_1, ..., \theta_k)$. The joint posterior distribution for all of the parameters is given by

$$p(\Theta|Data) = \frac{p(Data|\Theta)p(\Theta)}{p(Data)}$$

As in frequentist methods, the term $p(Data|\Theta)$ is the likelihood; while classical approaches stop at maximizing this value, the Bayesian approach relies on additional inputs. The term $p(\Theta)$ is referred to as the prior, representing information that the researcher has about the weight potential parameter values should be given. $p(Data)$ is a normalizing constant arrived at by marginalizing the probability of observing the data given the prior. For continuous distributions, this process relies on evaluating the integral

$$p(Data) = \int p(Data|\Theta)\,\mathrm{d}\Theta$$

For the vast majority of situations there is no analytical solution to this integral, and thus it requires a numeric approximation. This can be quite computationally intensive, especially in high-dimensional situations such as a multilevel model with many parameters and groups.

When $p(Data)$ it not known, the posterior can only be defined to a normalizing constant,

$$p(\Theta|Data) \propto p(Data|\Theta)p(\Theta).$$

For Markov Chain Monte Carlo methods, the specific value of $p(Data)$ is irrelevant to the estimation procedure because of the way MCMC methods sample from the posterior. As an example, the Metropolis-Hastings algorithm (Metropolis et al. 1953) is a Markov Chain Monte Carlo (MCMC) algorithm designed to estimate parameters by sampling directly from the unnormalized posterior $p(\Theta|Data)$ (Carlin and Louis 2008). MH uses accept/reject sampling to simulated values from the posterior, where the The ratio of proposed to current posterior values in the Markov chain only requires the unnormalized posterior since p(Data) cancels. Given a symmetric proposal density (i.e. the Metropolis algorithm), the accept/reject ratio α is defined by the ratio of the posteriors

$$\alpha = \frac{p(\Theta^{prop}|Data)}{p(\Theta^{curr}|Data)} = \frac{\frac{p(Data|\Theta^{prop})p(\Theta^{prop})}{p(Data)}}{\frac{p(Data|\Theta^{curr})p(\Theta^{curr})}{p(Data)}} = \frac{p(Data|\Theta^{prop})p(\Theta^{prop})}{p(Data|\Theta^{curr})p(\Theta^{curr})}$$

For many applications, Metropolis offers a simple yet powerful algorithm for simulating the posterior, but for high-dimensional models random-walk Metropolis can be computationally inefficient.

Bayesian methods are numerically equivalent to frequentist approaches given an uninformative prior extending across all possible values. However, it is usually the case that not all of these values are equivalently probable; for example, if a stock had a CAPM β on the order of 1,000,000, we would likely assume that something was wrong with our analysis. Even much smaller parameter values would give us pause, and incorporating this information through the prior distribution can both improve our estimates and make the integral much more manageable.

Objections to using prior information center around the subjective nature of the prior, but prior information can actually be necessary in small-data settings to mitigate false discoveries with noisy variables. More modern approaches to Bayesian modeling regard prior information as a form of regularization that provides necessary structural information about a model rather than the full state of prior knowledge (Gelman, Simpson, and Betancourt 2017). For example, if we were fitting a CAPM regression, we would expect the average relationship between members of a group (stock returns) and their average (the index) to be close to 1. While the prior always plays a role, in a setting with large amounts of data, the likelihood typically swamps the prior's initial weights for the credible values.

Advances in Markov Chain Monte Carlo

Because of the difficulty in evaluating high-dimensional integrals, multilevel models were computationally intractable for large numbers of groups. Recent advances Bayesian statistical computing now allow researchers to fit more complex, higher-dimensional models that were previously infeasible with standard MCMC techniques.

We estimate our model using Hybrid Monte Carlo, an algorithm originally developed for applications to statistical physics (Duane et al. 1987). This algorithm, now more popularly called Hamiltonian Monte Carlo (HMC) (MacKay 2003), uses accept/reject sampling of unnormalized posteriors similar to Metropolis. HMC therefore retains the flexibility of Metropolis in terms of model and prior selection. Unlike Metropolis, however, HMC leverages the gradient of the log posterior to guide MCMC samples. The resulting algorithm samples more efficiently than random-walk Metropolis (Neal 1996) without the restriction of fully specified conditional posteriors in Gibbs Sampling, and scales well in higher dimensions.

One of the major barriers to practical implementation of HMC is the necessity to tune a substantial number of parameters (Betancourt et al. 2017). Stan software (Carpenter et al. 2017) provides Bayesian practitioners with an implementation of HMC that automatically selects these parameters based on the particular application. Key components of Stan software include an automated gradient computation library (Carpenter et al. 2015) and the No U-Turn Sampler (NUTS) (Hoffman and Gelman 2014). The math library accurately computes gradients for any general model programmed in Stan, while NUTS ensures that the HMC samples from the log posterior in a computationally efficient manner.

Stan provides interfaces to a variety of programming languages used for statistical analysis, including R, Python, and Matlab. The R package BRMS (Bürkner and others 2017) used in this analysis offers a flexible implementation of Stan software designed specifically for Bayesian multilevel models. This package automatically translates multilevel models specified in R to Stan code, which is then compiled and run on the user's platform. The flexibility in prior and model selection offered by HMC provides academics and practitioners with the capability to specify target non-Gaussian generative distributions and a broad class of priors for reasons other than the ability to easily fit the integral of interest.

Multilevel Models for Pricing Equities

In this section we first discuss the relationship of the proposed multilevel models to existing models, demonstrating that they generalize approaches like Fama-French and Fama-Macbeth regression. We then talk about how to interpret these models and certain inferential advantages that the model structure contains. Finally, we discuss some of the benefits of the Bayesian nature of the model as it relates to model checking and prediction.

Generalizing Existing Models

For simplicity, let us start with estimating an asset's CAPM β via linear regression. Exposure to the market factor, β, is measured as the slope of a regression with the security's return R_i as the explained variable and the return of the market-cap weighted index X_{mkt} as the explanatory variable. The α component, or the intercept of this regression, is meant to represent the return of a security in excess of what can be explained through exposure to the market. The model is given by

$$R_i = \alpha_i + \beta_i X_{mkt} + \epsilon_i$$

Each security i is given it's own β_i and α_i for securities $1, ..., n$.

We might imagine a competing model, where there is only a global exposure to the market, and the securities did not have different slopes; this is common in other contexts, and is just a pooled regression. However, in the case where we have information about group-level heterogeneity it doesn't make sense to ignore or discared that information. We can think of that idea as similar in spirit to the "beta of 1" model in Elton, Gruber, and Urich (1978).

As discussed above, multilevel models generalize these cases, and directly condition the slope and intercept on both global (pooled) information and group-level information. If information is strong enough for either approach, the model will converge to global or group-level models. We can think of these as a set of related models, where R_i is related not only to stock-specific information (β_i and α_i) but also to global parameters (β_0 and α_0). This model is given by

$$R_i = \alpha_0 + \alpha_i + \beta_0 X_{mkt} + \beta_i X_{mkt} + \epsilon_0 + \epsilon_i$$

for stock i in stocks $1, ..., n$.

If the underlying groups are highly similar, it will approach the "pooled" model, where the between-group variance is set to 0. If they are very different, it will approach the "unpooled" model (the traditional approach

to β estimation), where the between-group variance is set to infinity. The extent to which the estimates are pooled is based directly on the ratio of the variance between groups and the variance within the group being estimated (Gelman and Hill 2006).

Another way to think of this is that stocks might be drawn from a meta-distribution; that is to say they have some characteristics in common (they are all equities), but also are distinct observations that makes them somewhat different (they are in different industries and have different business models). This approach is quite flexible, and could be extended by including group-level variation at the industry level in addition to the stock level, for example, or by including additional covariates. The normal way we estimate β is simply a special case of the multilevel model.

Multilevel models generalize this relationship by fitting both of these steps at once (Gelman and Hill 2006, 240, 270). Once we jointly estimate these steps, we can interpret α_0 as the return of the average stock and β_0 as the average relationship between returns and the factor X of interest (in this case a Market β). Another way to conceptualize this is to think of β_0 as the unconditional forecast for a stock that has no information. As such, these models attempt to directly answer the questions we are interested in. What is the unconditional relationship between systematic risk and return? Or in the case of testing whether a factor is priced, how would we price a new security based on the underlying factor of interest? Given a change in a stock's book-to-market ratio, say, how should we expect that to change its subsequent return?

Multilevel models also contain techniques like Fama-Macbeth or cross-sectional regression if, instead of stocks, we use each date as a group. To see how this works, let us imagine a multilevel model

$$R_d = \alpha_0 + \beta_0 X_{mkt} + \alpha_d + \beta_d X_{mkt} + \epsilon_0 + \epsilon_d$$

where d is an indicator for the date of measurement. Fama-Macbeth is typically measured with a cross-sectional regression measured at each date and subsequently averaged. This is the same as the regression above as σ_{β_d}, or the between-group variance of the market β parameter approaches ∞. The group-level coefficients are differences around a mean, and are forced to have mean 0, so if all observations are treated as independent, β_0 will simply contain the average across all the dates, while the variance between the dates will be entirely captured by β_d.

While the CAPM model is relatively simple, estimating market β properly is by no means a trivial problem. Currently financial economists rely on ad-hoc approaches to regularizing otherwise noisy estimates of market β, including estimating the correlation and variance component of market β separately and over different windows, arbitrarily choosing a shrinkage parameter (Frazzini and Pedersen 2014), and using a two-period measurement of β and adjusting based on the difference (Blume 1975). All of these estimators prevent the direct interpretation of the market factor as the slope of a regression line between a stock and a market factor. Instead, the multilevel model preserves this intuition and regularizes the estimates using a method that has well understood statistical properties. Therefore, an advantage of multilevel models of market β is that they provide statistical formalism in their approach to regularization that many historical approaches leave aside.

Multilevel models directly relate to an existing model for market β already common in financial economics proposed by Vasicek (1973). Vasicek argues for incorporating cross-sectional information in estimates of β, with the average β informing the estimates of the various stocks. This information is commonly referred to as a "prior" for how to adjust the aggregate stock β, with the empirical average of a set of regression loadings used in an empirical Bayes framework for adjusting the stock betas. This is implemented as a "two pass" model, which estimates a prior from the data and then uses it to inform our estimates in the data.

However, inferring a prior from observed data is effectively using the data twice, and thus can overstate the strength of an effect (Gelman, Simpson, and Betancourt 2017). Additionally, if we view the average market β as a parameter in the model (rather than an empirical prior), then the underlying security-level regressions derived from the first-pass regressions would suffer from omitted variable bias. If we don't include the pooled parameter in the individual regressions, those loadings are likely to overestimate the security-level parameters. Similarly, unmodeled differences between firms could bias the estimates of beta in a fully pooled model. In addition, the two-pass version of the model fails to account for potential correlation

between parameters (for example between the slope and intercept term), and it is unclear how we could extend the approach to incorporate additional parameters. We now have both the statistical framework and computational power to estimate these parameters jointly.

Inference for anomalies

While traditional tests of CAPM rely on unpooled models (or two-pass models as discussed above), attempts to identify generative factors in the sense of Arbitrage Pricing Theory often involve testing the returns of portfolios sorted on observable characteristics against a null hypothesis in order to identify factors that are related to future prices. Anomalous returns are thought to present evidence for a factor of interest, such as accounting variables or macro-economic sensitivities. However these tests are not sufficient to detect a true effect because they assume exchangeability of stocks within the market. If this assumption does not hold, it invalidates an attempt to attribute any causal meaning to the factor of interest.

A good example of how the traditional test can go awry is from Harvey (2017) where he discusses a new factor discovery:

> Here are the instructions that I gave my research assistant: (1) form portfolios based on the first, second, and third letters of the ticker symbol; (2) show results for 1926 to present and 1963 to present; (3) use a monthly, not daily, frequency; (4) rebalance portfolios monthly and once a year; (5) value weight and equally weight portfolios; (6) make a choice on delisting returns; and (7) find me the best long-short portfolio based on the maximum t-statistic.

While it is entirely possible that there were return correlations that could be mined in order to find a statistically significant "ticker symbol" effect, attributing any meaning to this seems foolish!

In tests of factor prices, we generally assign stocks to a portfolio based on some kind of sorting rule:

$$R_t = E(R_{i,t}|Q(BtoM_i) > 0.4)$$
$$R_c = E(R_{i,c}|Q(BtoM_i) < 0.4)$$

where $Q(x)$ is a function that returns a quantile of the variable of interest. The above is the rule that Fama and French (1993) use to estimate the return to the value factor. The difference between R_t and R_c is interpreted as a treatment effect of having high or low Book to Market (the return of a long-short portfolio).

However, this interpretation requires strong assumptions about assignment. To clarify this concern, we draw on the causal inference literature. Let us think about the portfolio sorting method as measuring the effect of a treatment t (e.g. high book to market) on R (a return). A portfolio assignment rule S assigns each stock to a portfolio that represents exposure to a factor. The treated group, usually a set of stocks that represent a quantile of a given fundamental characteristics, is then compared to a control group, a set of stocks in a lower quantile of that characteristic, weighted either equally or according to market capitalization. We typically measure the average treatment effect of being in a higher quantile $T = E(R_t - R_c)$ through $E(R_t) - E(R_c)$, where R_t is the return of the treated portfolio and R_c is the return of the control portfolio.

However, the difference between the two portfolios is only an estimator of the true treatment effect under strong assumptions. It is helpful here to bring back the assignment operator, S. It is not true, in general, that $E(R_t) = E(R_t|S = t)$, which is to say that the expected return to treatment is not the same thing as the expected return to treatment given the assignment of stocks to the treatment portfolio.

For this to be true, we need to assume that units of the treatment group are exchangeable with units of the control group. This is directly related to the notion of (conditional) independence. Even in a randomized setting, for inference between samples to be valid, it needs to be plausible that the assignment S of a stock into a portfolio needs to be independent of R_t and R_c and *all other variables over the unit* (Holland 1986).

Exchangeability is the idea that the the joint probability distribution of a series of random variables (e.g. $p(R_1, R_2, ..., R_n)$ for assets $1, ..., n$), the probability distribution remains the same for any and all permutations of the sequence. For some things this is believable, while for others it is not. For example, in the context of medicine it is common to control for race, age, and gender in studies by stratifying the treatment across the groups, and then comparing treated and controlled individuals within those groups. Similar exchangeability (or conditional independence) assumptions are the basis for "fixed effects" models in economics.

Claiming that the portfolio sorting method detects an effect that is caused by the observed characteristic makes strong and improbable exchangeability assumptions. Essentially it says that we can ignore the differences between different stocks cross-sectionally. It does not matter if Johnson and Johnson is in the treatment or the control group, because Apple is a good enough proxy for *what would have happened* if Johnson and Johnson had a different book to market. Since we do not know how to quantify all of the ways that the two securities are different, it is unlikely that we will be able to properly correct for all of the differences such that we can assign meaning to the effect.

Multilevel models based on repeated measurements of a time series use each unit as its own control. The exchangeability assumption is that Apple having a lower book to market in the past proxies for how Apple would be doing if it had a lower book to market now. There may still be reasons to be skeptical about whether results from long ago should be pooled together with more recent data, but this step helps to mitigate at least the most obvious violations. The model then looks at how similar the within-unit relationships are and pools them to the extent that they are similar, jointly estimating the average and group-level parameters.

Portfolio sorts, as opposed to regression, became fashionable tests of return factors because of concerns that it was inappropriate to pool the error terms of different securities in the same regression. This same logic is what motivates Fama-Macbeth regression, an approach to cross-sectional regression that recovers the mean return of a portfolio corrected for differences in the error term (Fama and MacBeth 1973). One of the major fears in financial econometrics is that if residuals are correlated across securities or time then standard errors are either over or understated dramatically. Portfolio sorts help to average securities and thus reduce the impact of correlated errors, and methods such as Fama-Macbeth allow for cross-sectional and time-series variation in errors and parameter estimates.

Multilevel modeling breaks apart the component of the error term that is stock-specific from the part that is correlated with other observations by construction, and thus does not suffer from the same concerns about whether errors are cross-sectionally correlated. Indeed, one of the ways to motivate multilevel models is as a large regression with correlated errors (Gelman and Hill 2006, 265). Errors are modeled both within each group (correlated at the firm level) and across groups. Were we to add additional levels to the model, we could accomodate industry or sector-level errors as well.

Model Checking and Prediction

Fully Bayesian models come with a posterior predictive distribution that is meant to describe the full generative model of the data used to construct them. As a consequence, posterior draws can be used to more accurately inform prediction, compare models, and evaluate model fit. Posterior distributions are not constrained to be symmetric around a point estimate; therefore making a buy or sell decision for a security does not have to just be based on symmetrical uncertainty around the mean. In a case where the posterior is particularly lumpy with a long tail, traditional confidence intervals may be difficult to calculate and misrepresent the true uncertainty.

We can also directly examine the impacts of targeting a given likelihood with our model by drawing samples from the posterior distribution of the trained model. Returns are often modeled as if they are log-normal even though this is known to be unlikely. The easiest way to identify which model better describes the data is through a variety of visual posterior predictive checks (Gabry et al. 2019). Using the returns of all the constituents in the S&P 500 between 2002 and 2007, we can check how good a CAPM regression is at modeling the returns by drawing from the trained posterior distribution. As an example, we fit the same model on stock returns data on a five year period starting in 2002 with two different likelihood targets: one

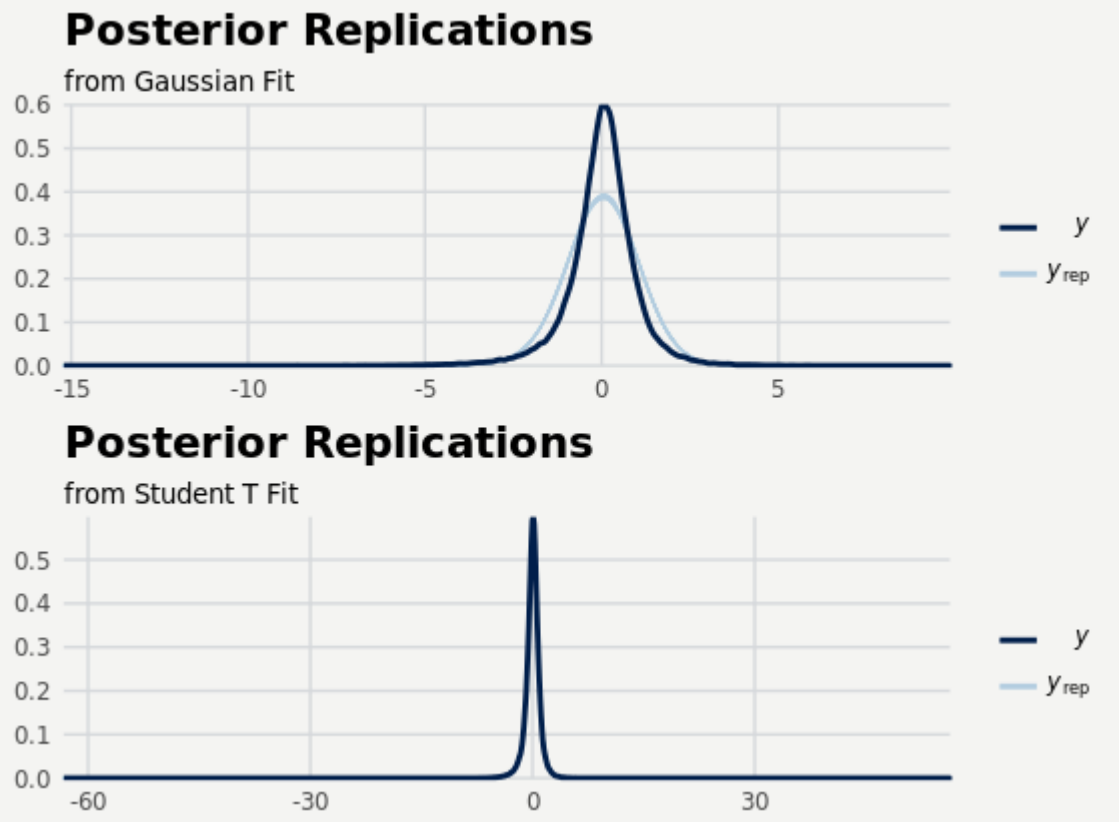


Figure 1: Posterior Density Comparisons: Log Returns Scaled to SD of 1

with a Gaussian likelihood and one with a Student T distribution. Returns are centered and scaled to a standard deviation of 1.

The light blue lines are posterior draws from the trained model (a multilevel one-factor market β model), while the thick dark blue line is a kernel density estimate of the observed data. We can see that the Gaussian model simply fails to fit the kurtosis of the distribution of stock returns. We chose to target a Student T distribution to better capture the tail behavior of stock returns, which substantially improves the model fit. We discuss the data, implementation, and results in the following section.

An application to risk and price forecasting

Implementation, Data, and Priors

To demonstrate the effectiveness of multilevel models in practice, we model the returns of the securities in the S&P 1500, going back to the index's inception in 1995. These data were chosen based on the ability to cover a wide swath of market capitalization and industries, and in part because of the availability of data to the researchers. While the CRSP database is typically used for these types of studies, we did not have access to CRSP during the research process. There may be advantages to using securities that are more likely to be liquid due to recent concerns about the impact of micro-cap stocks on factor estimates (Hou, Xue, and Zhang 2017), but either way it is not likely to have a large impact on the findings as this is a direct comparison of common methodologies on the same dataset. While results about factor prices may be different than using the merged CRSP/Compustat database, there is no reason in principle why the same models could not be applied to the more standard dataset. Either way, we make no strong claims about the

validity of specific factors, and leave those tests as an area for future research. For reference, as of February 21st, 2019 the S&P 1500 covers 91% of the market capitalization of the U.S. market and contains 1,506 constituents, with the smallest being $92 million.

In order to test multiple regression factor models based on stock fundamentals, we merge the stock returns data with the CIQ fundamentals database, using the most recent stated data available prior to the observed return. CIQ filings data are provided on an as-was basis, and as a consequence we can construct fundamentals data based on the information available to investors at the time, which helps prevent any look-back bias in the model. The explanatory variables chosen in the price forecasting model are the Z-score of a given signal for a given security at time t-1 relative to the cross-section. All returns are in excess of the risk-free rate, measured as the return on a 1 month US Treasury bill.

It is both helpful and necessary to choose assign priors to the parameters in our models. For the purposes of this paper we chose priors to regularize estimates rather than a highly informative prior on a pre-defined value. It would be unfortunate if the priors dominated the data in the analysis; to avoid this we use prior distributions that incorporate a all possible parameter values while putting a larger share of the density around a central tendency. While these priors do serve to inform the model, they are chosen to appropriately regularize estimates rather than for their ability to be computationally tractable. In the sense that they help to mitigate extreme parameter estimates, putting more prior density around zero (or 1 in the case of CAPM β) is more conservative than a uniform prior stretching from $-\infty$ to ∞, and serves the notion that ex-ante we do not believe there to be any effects. In a "small data" setting with (e.g. where we only have 36 months of data for a security), it can serve the important purpose of regularizing the information towards a value with better predictive properties. For detail on the priors used to fit the various models, please refer to the appendix.

Modeling Risk

Unregularized estimates of market β are well-known to be extraordinarily noisy. There are several approaches to correcting the estimates of β, but they are typically more rooted in the practicality of their results than any sort of statistical motivation. This class of estimators includes changing the window used to estimate the covariance and the correlation components of β (Frazzini and Pedersen 2014) and/or applying a shrinkage factor to the estimate by assuming that the estimate will revert to the grand mean (e.g. Blume (1975) and Vasicek (1973)). A distinct advantage of using a multilevel model is that the beta estimates will be shrunk towards the global mean *to the extent it is warranted*. This is to say that multilevel models, because they include both the fully pooled and the unpooled models within them, will eventually converge to either model should there be enough evidence. Without enough evidence, they fall somewhere in the middle.

Interestingly, there is historical precedent that advocates for the spirit of this approach. Black et al. (1972) argue that

> We would like to design a test that allows us to aggregate the data on a large number of securities in an efficient manner. If the estimates of the [stock αs] were independent with normally distributed residuals, we could proceed along the lines outlined by Jensen (1968) and compare the frequency distributions of the "t" values for the intercepts with the theoretical distribution. However, the fact that the [errors] are not cross-sectionally independent... makes this procedure much more difficult.

As discussed earlier, multilevel models are designed to share information across groups while allowing for correlations in the error term.

A common extension of the one-factor contemporaneous asset pricing model is the Fama-French 3 factor model, which uses portfolios sorted on book to price and company market cap to add additional information to the risk model. We test both of these models within the multi-level model framework and compare them to their unpooled counterparts and to each other, both on an in and out-of-sample basis.

The multilevel models we are interested in are given by

$$R_i = \alpha_0 + \alpha_i + \beta_0 X_{mkt} + \beta_i X_{mkt} + \epsilon_i + \epsilon_0$$

and

$$R_i = \alpha_0 + \alpha_i + \beta_1 X_{mkt} + \beta_2 X_{hml} + \beta_3 X_{smb} + \beta_{i,1} X_{mkt} + \beta_{i,2} X_{hml} + \beta_{i,3} X_{smb} + \epsilon_i + \epsilon_0$$

for security i in securities $1, ..., n$.

The no pooling models are given by

$$R_i = \alpha_i + \beta_{i,1} X_{mkt} + \epsilon_i$$
$$R_i = \alpha_i + \beta_{i,1} X_{mkt} + \beta_{i,2} X_{hml} + \beta_{i,3} X_{smb} + \epsilon_i$$

We estimate these models on four separate five-year periods and evaluate them by taking their parameter estimates and measuring their goodness-of-fit on the subsequent out of sample three year period. Typically we would do this on a rolling basis, as these estimates might get stale or not take advantage of new information, but the number of independent periods is still highly limited in the sample and the models are computational intensive to fit. The period in question is the full history of the S&P 1500 index through the year 2018, a 23 year period.

We can see that the multilevel models dominate their unpooled counterparts in every single out of sample period. Out of sample, the new approach has consistently higher R^2 and better measures of absolute predictive fit than the equivalent unpooled models. Surprisingly, the simple multilevel one-factor beta model actually does better than the unpooled three factor model in all but one period! The partially pooled three factor model dominates across the board.

Table 1: Out of Sample Model Performance using Posterior Median Prediction

Period Ending	Fit	Model	MSE	RMSE	Rsq
2003-12-31	No Pooling	3 Factor Model	1.11	0.72	16%
2003-12-31	Partial Pooling	3 Factor Model	0.95	0.67	24%
2003-12-31	No Pooling	CAPM	1.09	0.72	14%
2003-12-31	Partial Pooling	CAPM	1.04	0.70	18%
2008-12-31	No Pooling	3 Factor Model	0.96	0.63	12%
2008-12-31	Partial Pooling	3 Factor Model	0.84	0.59	18%
2008-12-31	No Pooling	CAPM	0.90	0.61	14%
2008-12-31	Partial Pooling	CAPM	0.86	0.60	17%
2013-12-31	No Pooling	3 Factor Model	0.45	0.47	20%
2013-12-31	Partial Pooling	3 Factor Model	0.42	0.45	24%
2013-12-31	No Pooling	CAPM	0.44	0.46	20%
2013-12-31	Partial Pooling	CAPM	0.42	0.45	23%
2018-12-31	No Pooling	3 Factor Model	0.56	0.50	21%
2018-12-31	Partial Pooling	3 Factor Model	0.54	0.49	23%
2018-12-31	No Pooling	CAPM	0.60	0.52	15%
2018-12-31	Partial Pooling	CAPM	0.58	0.51	19%

We can see why this is happening when we look at the dispersion of the beta and alpha estimates of the different models. Naive beta estimates can be extrordinarily large or small. Indeed even using a five year estimation window, we can see outlandish beta estimates like -1 or 3; the full range of the beta coefficients in the no-pooling model is -2.6 to 4.7. In contrast, the partially pooled model has much more reasonable

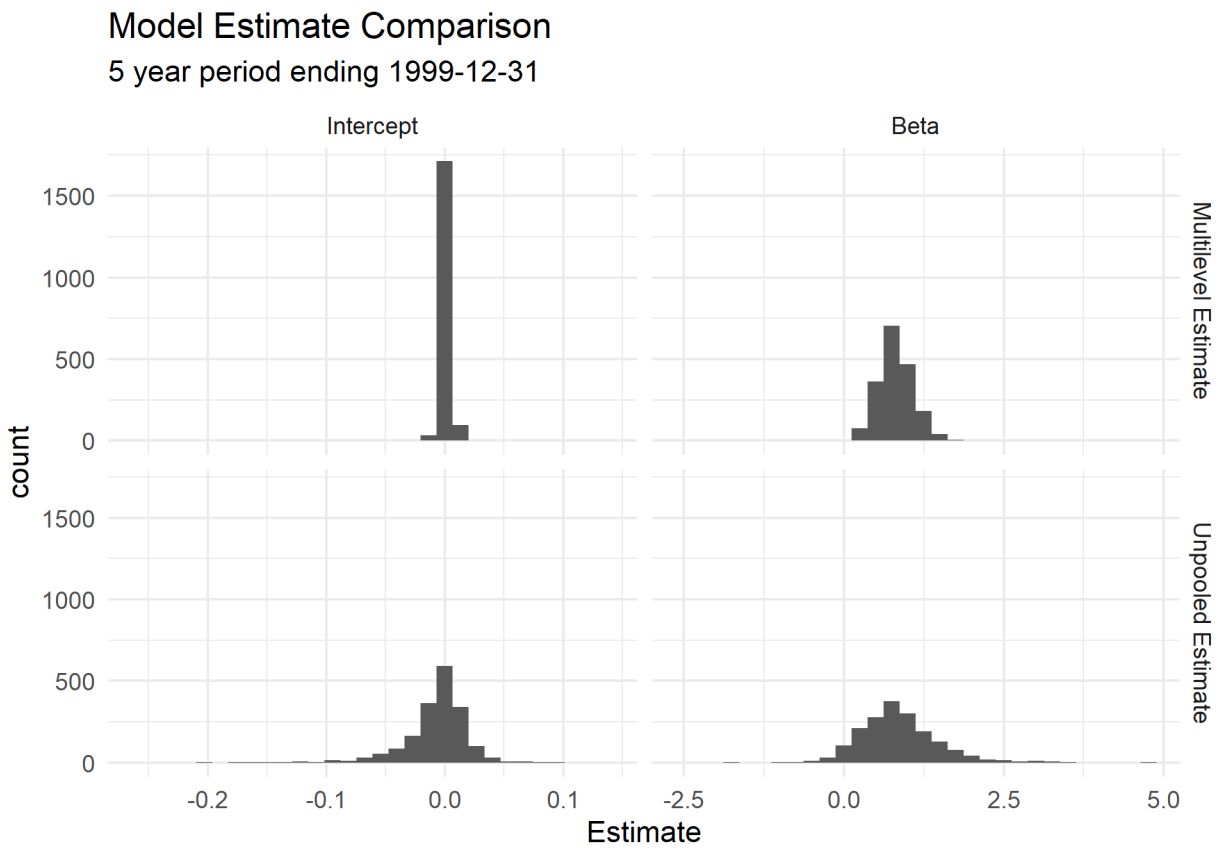


Figure 2: Stock Level Parameter Estimates for Multilevel and Unpooled CAPM Models

estimates for both the intercept and beta terms, ranging from 0.1 to 2. This is considerable evidence for the power of the market factor, which (when estimated this way) has a loading of close to 1 in every period. In other words, were we to see a new stock, our best bet is that it would have a market β of 1. The four five-year models have loadings on the market factor that range from 0.86 to 1.13, and the full-sample estimate is 1.06.

Given the popularity of the three-factor model (and similar risk models including additional factors like profitability, investment, and momentum), the substantial improvement in out-of-sample performance of the model is quite valuable. After all, the risk adjusted return left over after accounting for the market, the value effect, etc., is treated as the "skill" of a manager in a portfolio setting or the presence of a risk factor that is not priced by the market. The fact that regularization improves the aggregate fit of the model implies that the traditional least-squares estimate is overfitting the risk exposures in sample. The stabilized loadings of a multilevel model may be more appropriate for these tests given that they more accurately represent the generative process.

While this demonstration uses relatively simple risk models compared to, say, industry standards like Barra, they could easily be extended to include more parameters and additional levels. For example, one could imagine additional groupings at the country, industry, or sector levels. The loadings that come from that type of fit are interesting in and of themselves; for example, they would provide direct evidence for how the significance of well-known effects vary by sector.

The substantial improvement in estimating the exposure of a security to a factor is also valuable to people constructing portfolios based on quantitative signals. It is common to take a known signal, construct a factor portfolio, and then score securities based on their exposure to that portfolio. These scores are then used to construct an optimized portfolio. Because these models better estimate a security's score, they could generate a substantial advantage for quantitative asset managers.

Forecasting Prices

Because multi-level models directly model the cross-sectional/time-series component of stock returns, there is no need to sort stocks into portfolios to perform tests of return anomalies. Instead, if we imagine testing the 3 factor model of Fama and French (1993) within this framework, we would measure the exposure of stocks to changes in book to market value and market capitalization in addition to their aggregate beta. Stocks would have their own slopes and intercepts for each of these factors, but there would also be unconditional pooled estimates at the higher level of the model. These estimates represent the average relationship between one of these variables and return, and thus we would have a direct estimate of the average risk premium in the cross-section.

There are several advantages to this representation vs. traditional portfolio sorts. We remove some "degrees of freedom" that the researcher normally has when choosing how to sort a portfolio (Gelman and Loken 2013). As discussed above, this model does not treat all stocks as exchangeable but instead uses a stock as its own control. If I sort stocks into a portfolio and then use their average returns to estimate the returns to an anomaly, then I am making the statement that we should regard the stock returns on a given date used to form both treatment and control portfolios as exchangeable. This is inappropriate because we *know* that these stocks are not all the same.

Fitting this model allows us to directly test how good these factors are at forecasting security returns, while sorted portfolios simply see if there is a relationship between a lagged, observed variable and subsequent abnormal returns. Given a portfolio sorting rule, there is no clear way to do cross-validation or out of sample forecasting tests with that information, which severely limits the ability to test the specification for robustness or compare between models. The justification for these non-parmetric models of returns stem from fears of improperly pooling regression results across stocks with heterogeneity in their error terms. Because the multilevel model specifies separate independent and pooled error terms for each security, this is not a concern, and the stock-level variances are allowed to differ. Finally, the regularization inherent in the multilevel model structure helps to mitigate the effects of multiple comparisons.

Since Fama and French (1993) it has been common practice to use the HML and SMB factors (or some variant thereof) to proxy for underlying drivers of stock returns. This is because of the "dubious empirical content" of the CAPM model (Huberman 1982), which assumes that long-run returns are solely a function of systematic risk. Keeping in mind that because of differences in data and the choice of sorting date these portfolios will not match the HML and SMB portfolios from Ken French's data library, we can still construct equivalent portfolios from our own data and look at the explanatory power of the portfolios relative to a multilevel model. We use these factors as a proof of concept for the modeling approach simply because they are so well known within academia and the finance industry, not because they necessarily represent the best way to model securities. The model can be easily extended to incorporate other factors.

One note in interpreting these coefficients is that the market variable is centered to improve convergence to the appropriate posterior. Non-centered variables impose a correlation structure between the slopes and the intercept term, and impede the speed of MCMC convergence. As a consequence the intercept term in the model contains the average unconditional response of the Y variable, rather than the traditional "risk adjusted return." If we were to subtract the mean market return over the period from the intercept term, we would recover the traditional estimate of α.

The average coefficient on the changes in the underlying generative variable will determine whether, on average, that relationship is priced in the market. In other words, does a single unit change in market cap lead to a change in returns in the next period. In this sense it can be thought of in as similar to a Fama-Macbeth approach, which attempts to determine the price of a factor given its exposure to a portfolio sorted on a characteristic or a t-test on a sorted set of stock portfolios, but the relationships cannot be directly compared given that Fama-Macbeth estimes are conditional on a portfolio sorting rule.[1]

The parameter estimates of the multilevel model are more directly interpretable risk premia than the returns of sorted portfolios, as they represent the return in period $t+1$ given a unit change in variable x_t. In the case

[1]For a comparison of the multilevel model loadings relative to other common models like t tests on portfolio returns and Fama-Macbeth regression, please refer to appendix A.

of this model the lagged book to market and size variables are scaled, and so the parameter values represent the conditional return in the following month given a standard deviation change in size or book to market (relative to the average in the cross-section at a given date). The estimates for the intercept, size, and book to market varaibles displayed in percentage terms for interpretability.

For a given change in the standard deviation of log market cap (approximately a 3.5x change in size), we can expect returns to change somewhere between -3 and 4 basis points 95% of the time. A change of about 60 in the book value per share of a firm should lower returns, on average, by about 32 basis points in the following month. Each firm has different slopes in the model, and thus some might be more or less sensitive to similar changes, but if we had no conditional information about the firm (say a new stock listing), this is how those variables should impact the prices. This should largely line up with our intuition about how size and value-based strategies have performed over the last 20 years or so; if we were to use this model to truly search for meaningful drivers of return we would likely want to increase the historical sample size.

Table 2: Multilevel Parameter Estimates over Full Sample

Variable	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept	0.73	0.02	0.70	0.77	3714	1
LagBtoM	-0.32	0.03	-0.39	-0.25	2355	1
LagLogMktCap	0.01	0.02	-0.03	0.04	3715	1
CenteredMkt	1.06	0.01	1.04	1.08	2699	1

In order to understand how the various periods were driving the aggregate parameter values, we fit the model on the same five-year sub-periods as we did the risk models. We see the most consistency in the market factor, which circles around a β of 1. While the value factor (lagged Book to Price) does have a consistently negative loading over the full period, the technology run-up of the late 1990s represented the most punishing period for the variable, when a standard deviation change in Book to Price was associated with a 1 to 2% lower return in the following month! Size does not seem to have a consistent relationship across the sample, which makes sense given that its full-sample confidence interval contains 0.

We can also estimate the extent to which these groups are similar or varying by using the "pooling factor," which represents the degree to which stock-level estimates are pooled towards a central value. A pooling factor (denoted λ) close to 1 implies that the groups are highly similar (the between-group variance is next to 0), while a λ close to 0 implies the opposite. For a simple multilevel model with a group-level intercept term α, the pooling factor is given by

$$\lambda = 1 - \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_y^2}$$

and, in general, it is given by

$$\lambda = 1 - \frac{V_{k=1}^k E(\epsilon_k)}{E(V_{k=1}^k \epsilon_k)}$$

where ϵ_k is the error term for group k and E represents the posterior mean (Gelman and Pardoe 2006). The intuition behind this measure is that the if we compare the variance between-groups to the total variance in the model, we can estimate to what extent the overall variance is driven by differences between securities. We use draws from the posterior distribution to calculate the variance in the parameter estimates, and thus the level of pooling.

We should be interested in this measure because if a fundamental characteristic has a highly-varying relationship in the cross-section it still might be useful in forecasting, but operates differently from how we typically think of latent linear pricing factors. A parameter that claims to be a strong latent linear pricing factor

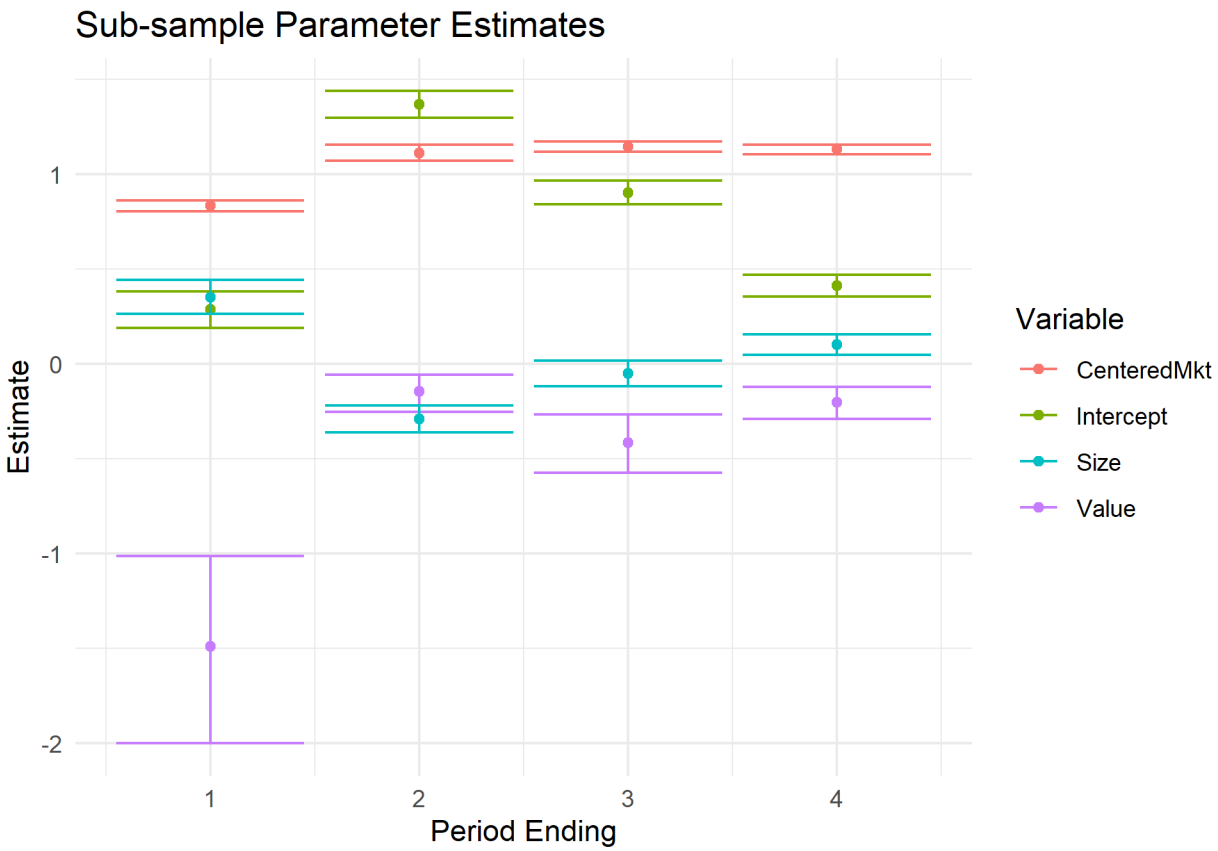


Figure 3: Sub-sample Estimates of Parameter Values

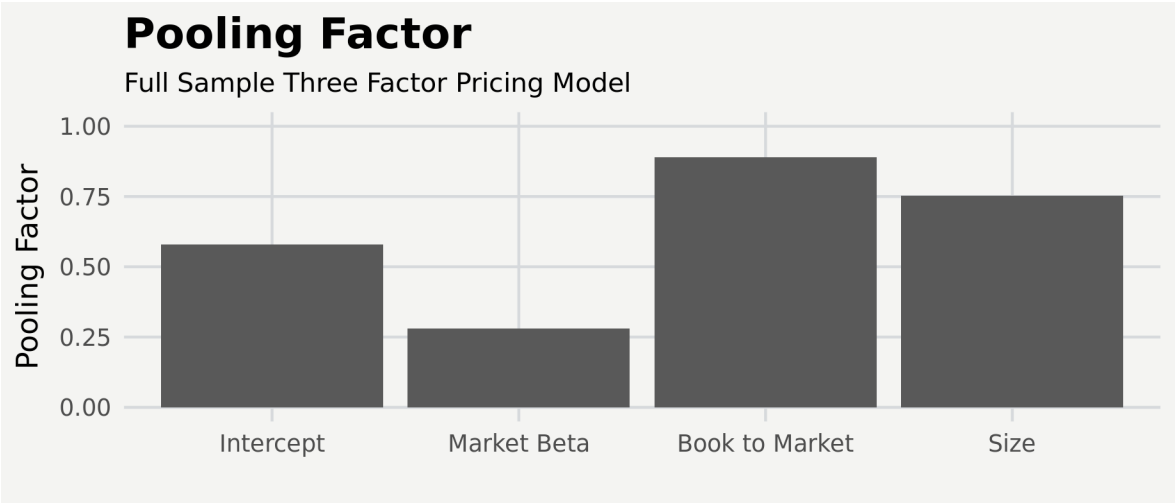


Figure 4: Pooling factors for full sample model

should have a highly similar relationship across all securities. By contrast, Market β is a risk sensitivity that should vary across securities.

These pooling factors estimated over the full sample tell us that the book to market and size parameters have relatively stable relationships, while the market beta parameter varies the most cross-sectionally. This lines up with our general beliefs, and makes sense given that adding book to market and size factors in the multilevel risk model did help the model.

multilevel models also provide results that are robust to choices in the sorting rule, and the return forecasts baked into the posterior interval contain information about both the magnitude and certainty of the relationship. As a consequence, a forecaster who has fit one of these models could conceivably optimize them relative to any portfolio construction rule. In addition, someone using a multilevel model can take into account estimation uncertainty in the model by using the full posterior predictive distribution, which can substantially impact optimal portfolio choice (Kan and Zhou 2007; Klein and Bawa 1976). Because the model does not assume a specific direction of an effect that applies to all securities, an investor could allocate positions accounting for the fact that securities have relationships to the underlying variables that are fundamentally different instead of relying on the depth of exposure to a given sorted portfolio conditioning on that portfolio having a significant return.

As an example, if we wanted to test the return of a market-cap weighted portfolio given the fitted model, we would simply predict the returns of each security from the full posterior distribution, weighting the predicted values according to the relative market cap of the securities in question. This has direct parallels to the approach political scientists use to stratify regression results to expected demographic turnout of their election models (Ghitza and Gelman 2013), where the sensitivity of the parameters is constructed with a multilevel regression model and the estimates are then weighted based on expected (realized) turnout in the coming (previous) election.

New Forms of Testing and Inference

As with anything, a new approach to modeling allows us to ask questions that were previously difficult with previous frameworks. For example, because unpooled models do not share any information across securities, we cannot use them to make predictions about securities not included in the model. By contrast, a multilevel model can be used for kfold or leave one out cross-validation, a direct test of the ability of a model to price a held out asset.

Currently academics test factor models based on the return left over after regressing the security's or portfolio's return stream on the relevant long-short portfolios. With a multilevel model, we can directly figure out how much information the model can learn from all securities but one, and use that information to price the other one. Better models should be able to price returns in-sample more accurately, and as these sorts of tests are independent of any notion of the stastical significance of a given factor, they can help to mitigate the multiple testing problem. Rather than concerning ourselves with the fact that we can reject a null hypothesis at a given confidence level, we can ask whether the information included in the factor actually helps us accurately price securities. While we might still be interested in traditional posterior or confidence intervals within a model, this type of test can complement that type of study by comparing the new model's expected predictive density to the old.

Because these models are computationally intensive to fit (on the order of 2 or 3 days for a large sample), we rely on the approach of Vehtari, Gelman, and Gabry (2017), who develop an approximate criteria for the pointwise predictive density of a model fit via MCMC techniques. By leveraging the existing MCMC draws within the model, we can approximate exact leave-one-out cross-validation (LOO) by using importance weights. The right tail of the importance weights are fit to a Pareto distribution in order to prevent extraordinarily large importance weight values, and after calculating testing to make sure that the shape parameter of the distribution is below the threshold such that the central limit theorem holds.

In other words, if we fit a model predicting y given parameters θ, we generally are claiming that the data contains information about the relationship of θ and y such that we could predict certain things about y

given a new dataset of *theta*. If we observe n new data points, we are interested in how well we expect to predict those data points. Given a posterior distribution $p(y|\theta)$ and a prior $p(\theta)$, we also have a posterior *predictive* distribution, or information about the expected distribution of a new dataset $\tilde{y}$, $p(\tilde{y}|y)$. Vehtari, Gelman, and Gabry (2017) define a measure of expected pointwise predictive density for a new dataset as

$$\Sigma_i^n \int p_t(\tilde{y_i}) \log p(\tilde{y_i}|y) \mathrm{d}\tilde{y_i}$$

where $p_t(\tilde{y_i})$ is the distribution of the true data-generating process for the new dataset. Since we do not know this value, we have to approximate it with cross-validation.

The expected log predictive density obtained via LOO is the sum of the log probability of all possible points y_i given the rest of the observed dataset. The way we can approximate these values is by constructing importance samples from the posterior samples. Importance ratios for observations that are conditionally independent are given by

$$\frac{1}{p(y_i|\theta^s)},$$

where θ^s are s samples from the posterior of the trained model. Because the variance of these values can be infinite, we smooth them by fitting a pareto distribution to the right tail of the importance ratios, making sure that the shape of the fitted distribution for each datapoint does not exceed values that would violate the central limit theorem.

From this procedure, implemented in the R packages loo (Vehtari et al. 2019), rstanarm (Goodrich et al. 2018), and BRMS (Bürkner and others 2017), we can compare the expected log predictive density of two different models trained on the same data. This value (and its standard error) provides a direct way to compare the expected predictive power of the two models given the sample. We find a considerable advantage for the 3 factor model in terms of expected log predictive density; the difference in expected errors is 67 standard errors away from 0.

Model	Difference in ELPD	Standard Error
CAPM - Fama French 3	-13027.36	208.0151

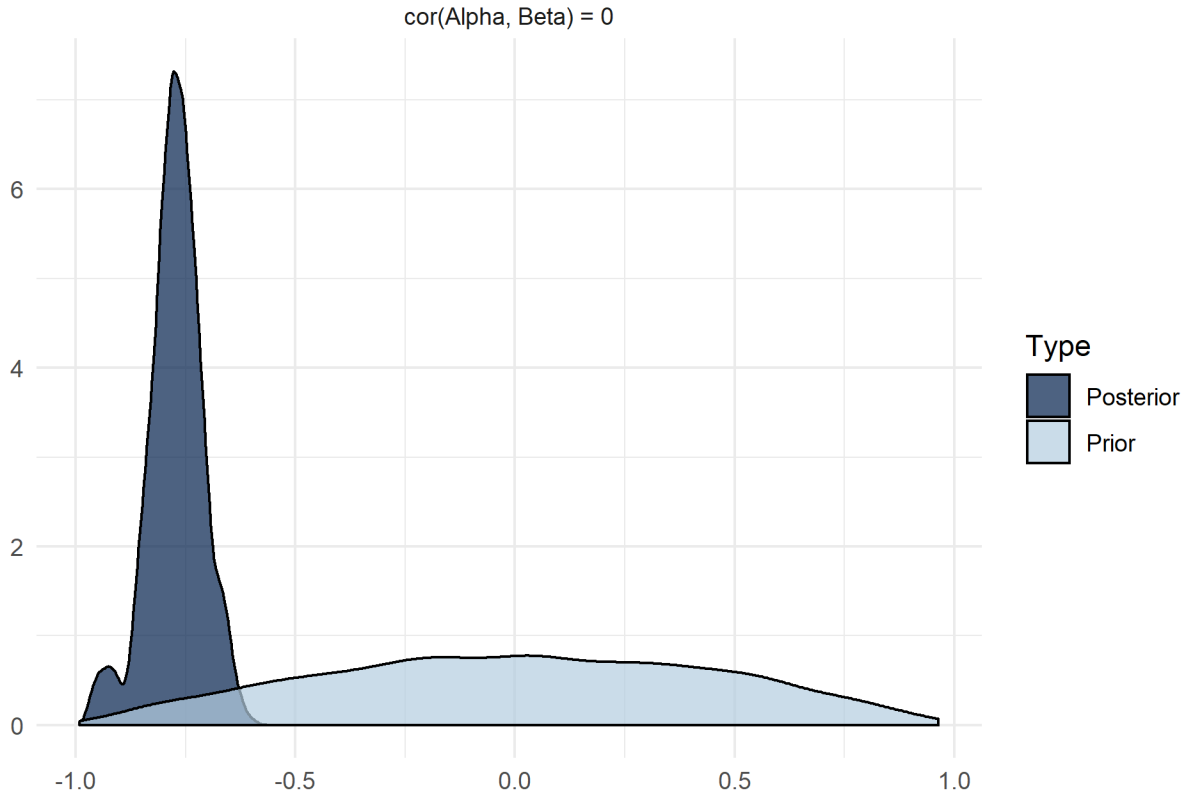
Additionally, Bayesian models are fully generative, which allows us to directly test more complicated hypothesis. Recently there has been a revived interest in the idea that market β might not be fully compensated, and as a consequence there are better risk-adjusted returns than expected to holding stocks with a low exposure to systematic risk. Essentially, the idea that we want to test is that stocks with lower β have higher α, and thus there should be a negative cross-sectional correlation between the two.

Table 4: '*': The expected value under the hypothesis lies outside the 99.9%-CI. Posterior probabilities of point hypotheses assume equal prior probabilities.

Hypothesis	Estimate	Est.Error	CI.Lower	CI.Upper	Evid.Ratio	Post.Prob	Star
cor(Alpha, Beta) = 0	-0.77	0.06	-0.97	-0.6	0	0	*

At least in this sample, and in this dataset, we find evidence that the security market line is flatter than we might expect, even after appropriately regularizing estimates of β. Importantly, this result holds even though the parameters are scaled and centered. If the parameters were not centered, there would be a structural negative correlation between the slope and intercept term, which is true whenever both the dependent and independent variable are positive on average. However, this fairly strong correlation even holds after we

center the market and average security return.



Discussion

Multilevel models offer promising properties for forecasters and researchers alike. The regularization inherent in the structure of the model makes multiple comparisons less problematic, all the while improving out of sample forecasts relative to OLS regression models. The aggregate, unconditional parameter estimates have superior inferential properties because of the structure of the model. As opposed to portfolio sorts, these models are within-unit comparisons that are aggregated to the extent warranted by the structure of the data, which allows for an easy interpretation of their parameter values (e.g. the marginal impact of Y given a one unit change in X).

Bayesian models also quantify a full posterior distribution, and a fully Bayesian multilevel model allows the uncertainty associated with a signal to vary at the stock level. This has substantial implications for optimal portfolio choice rules in the sense that the models produce a forecast with quantified uncertainty that could be optimized relative to a utility function or decision rule (Kan and Zhou 2007). Traditional point-estimate based forecasts are unable to take into account varying uncertainty for a signal, and only have an associated standard error term on which to make decisions. In the case of non-normal posterior distributions, a confidence interval based on a Gaussian distribution may be an inappropriate approximation of uncertainty. These qualities make the models highly desirable for practitioners who have to make portfolio implementation decisions with respect to a mean-variance optimal decision rule and academics who want to fully quantify the uncertainty of their models.

Hamiltonian monte carlo methods allow researchers to target more complicated likelihoods, specify better priors, and fit high-dimensional models without having to worry about the tractability of the posterior distribution. For anyone excited to use these models, the one caveat that must be mentioned is that fitting

multilevel models can take a very long time. Markov Chain Monte Carlo techniques offer the ability to perform integrals where there is no analytical solution, but they can take a long time to converge in high-dimensional spaces. For quicker inference, it may be appropriate to use other techniques for approximating integrals such as variational bayes, which uses latent gaussian distributions to approximate the posterior. In some cases it may also be possible to use more traditional optimization methods for approximate inference that use maximum likelihood or restricted maximum likelihood criteria (such as those implemented in lme4).

All in all, multilevel models offer a solution to many of the problems currently facing the areas of asset pricing research focused on discovering factors that are linearly related to prices. These models have advantages over traditional linear regression models because of their ability to aggregate within-unit treatments, their regularization properties, and improved forecasting ability. While at first the proposition may seem relatively alien to the finance field, practitioners and researchers should take comfort in the fact that many of the regression models they already use are simply specific cases of this general approach to modeling.

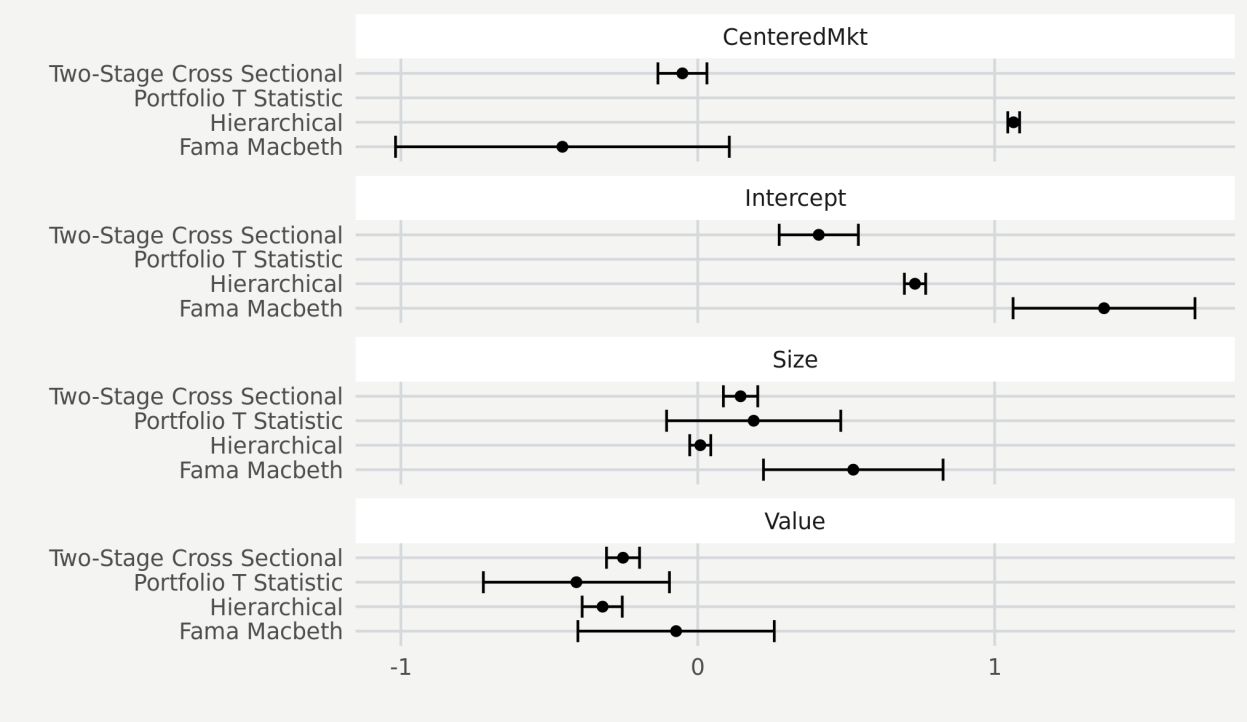


Figure 5: Full Sample Estimates Compared to Existing Techniques

Appendix A: Additional Graphs and Statistics

While the coefficients are not directly comparable (one is a response to a change in a characteristic while the other is a change in portfolio return), we compare the full sample and subsample fits of the multilevel model to a two-stage cross-sectional regression, portfolio t-tests, and fama-macbeth regressions. Figures 5 and 6 conntain these comparisons.

Appendix B: Prior Choices

All models

Across models we fit the β_{mkt} parameter with a T-distributed prior centered on 1 and a scale parameter of 1. This is centered on 1 to represent the expectation that the average stock will have a risk exposure that roughly moves with average dollar invested in equity, but weak enough that it does not rule out any estimate far away from 1. I put a prior of half-Cauchy(0,2) on the group-level standard deviations, and an lkj(2) prior on the correlation matrix for the group-level effects (named for Lewandowski, Kurowicka, and Joe (2009)), which puts more probability mass in regions more similar to the identity matrix.

Risk models

We put a Normal(0,1) parameter on the intercept term and all β's other than market. Stock returns (the dependent variable) are scaled and centered, and because both the dependent and independent variables are centered this value represents the average response when the market is at its average value. In other words, when the market has its average return, how much more or less does a stock return versus its average.

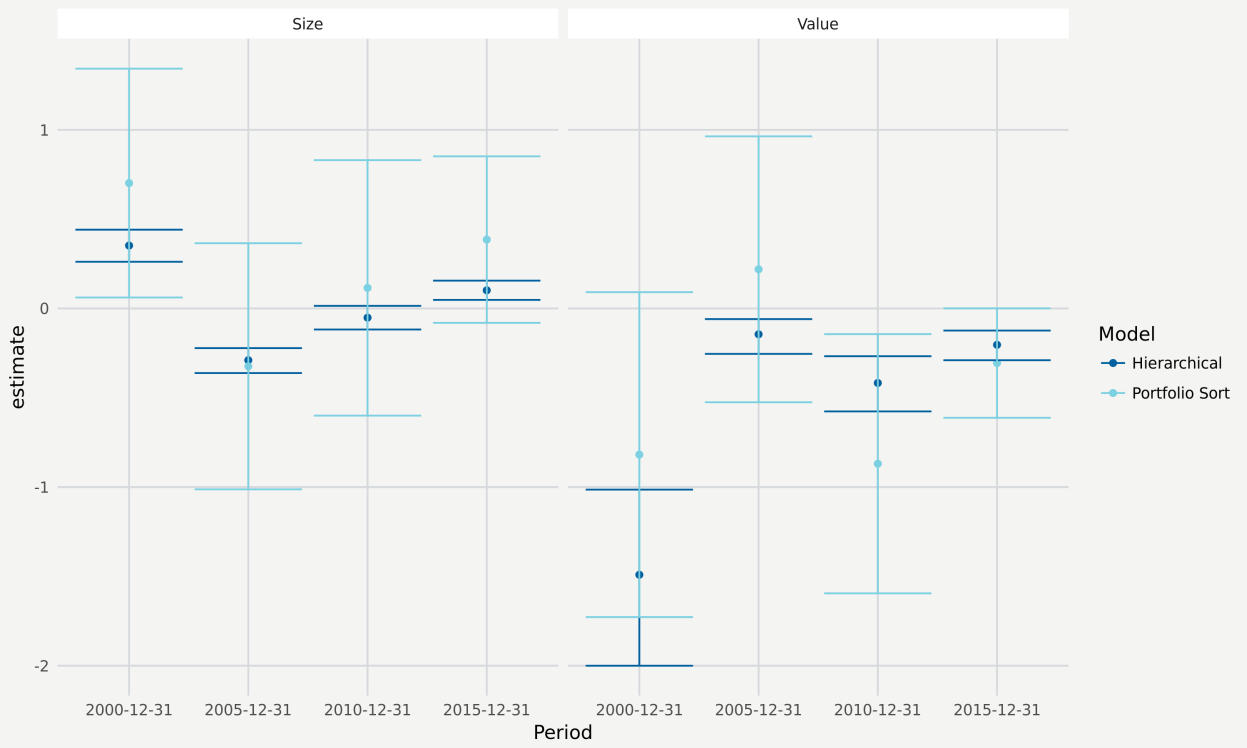


Figure 6: Subsample Estimates Compared to Portfolio T-Tests

Placing that in the range of a standard deviation of the stock's return is not particularly restrictive, but it does help to put less weight on extreme posterior values.

Factor pricing models

This model centered the return-based parameters but did not scale them. We used Normal(0,0.05) prior for the intercept, implying that about 68% of the probability mass for the intercept parameter would be in the [-5%, 5%] (the average reponse was 0% by construction). For the non-market factors we used a Normal(0,1) prior.

Summary

None of the priors on the coefficients ended up being highly informative; the posterior variance is far lower than the prior variance. As an example, the estimate for LagLogMktCap was 0.01 with a CI of [-0.03,0.04], while the prior standard deviation was 1.

References

Bates, Douglas, Deepayan Sarkar, Maintainer Douglas Bates, and L Matrix. 2007. "The Lme4 Package." *R Package Version* 2 (1): 74.

Betancourt, Michael. 2017. "A Conceptual Introduction to Hamiltonian Monte Carlo." *arXiv:1701.02434 [Stat]*, January. http://arxiv.org/abs/1701.02434.

Betancourt, Michael, Simon Byrne, Sam Livingstone, and Mark Girolami. 2017. "The Geometric Foundations of Hamiltonian Monte Carlo." *Bernoulli* 23 (4): 2257–98. https://doi.org/10.3150/16-BEJ810.

Black, Fischer, Michael C Jensen, Myron Scholes, and others. 1972. "The Capital Asset Pricing Model: Some Empirical Tests." *Studies in the Theory of Capital Markets* 81 (3): 79–121.

Blume, Marshall E. 1975. "Betas and Their Regression Tendencies." *The Journal of Finance* 30 (3): 785–95.

Bürkner, Paul-Christian, and others. 2017. "Brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28.

Carlin, Bradley P., and Thomas A. Louis. 2008. *Bayesian Methods for Data Analysis, Third Edition*. CRC Press.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 26 (1). https://doi.org/10.18637/jss.v076.i01.

Carpenter, Bob, Matthew D. Hoffman, Marcus Brubaker, Daniel Lee, Peter Li, and Michael Betancourt. 2015. "The Stan Math Library: Reverse-Mode Automatic Differentiation in C++." *arXiv:1509.07164 [Cs]*, September. http://arxiv.org/abs/1509.07164.

Duane, Simon, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. 1987. "Hybrid Monte Carlo." *Physics Letters B* 195 (2): 216–22. https://doi.org/10.1016/0370-2693(87)91197-X.

Elton, Edwin J, Martin J Gruber, and Thomas J Urich. 1978. "Are Betas Best?" *The Journal of Finance* 33 (5): 1375–84.

Fama, Eugene F, and Kenneth R French. 1993. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics* 33 (1): 3–56.

Fama, Eugene F, and James D MacBeth. 1973. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy* 81 (3): 607–36.

Frazzini, Andrea, and Lasse Heje Pedersen. 2014. "Betting Against Beta." *Journal of Financial Economics* 111 (1): 1–25.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 182 (2): 389–402.

Gelman, Andrew, and Jennifer Hill. 2006. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.

Gelman, Andrew, Jennifer Hill, and Masanao Yajima. 2012. "Why We (Usually) Don't Have to Worry About Multiple Comparisons." *Journal of Research on Educational Effectiveness* 5 (2): 189–211.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'P-Hacking' and the Research Hypothesis Was Posited Ahead of Time." *Department of Statistics, Columbia University*.

Gelman, Andrew, and Iain Pardoe. 2006. "Bayesian Measures of Explained Variance and Pooling in Multilevel (Hierarchical) Models." *Technometrics* 48 (2): 241–51.

Gelman, Andrew, Daniel Simpson, and Michael Betancourt. 2017. "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy* 19 (10): 555.

Ghitza, Yair, and Andrew Gelman. 2013. "Deep Interactions with Mrp: Election Turnout and Voting Patterns Among Small Electoral Subgroups." *American Journal of Political Science* 57 (3): 762–76.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2018. "Rstanarm: Bayesian Applied Regression Modeling via Stan." http://mc-stan.org/.

Harvey, Campbell R. 2017. "Presidential Address: The Scientific Outlook in Financial Economics." *The Journal of Finance* 72 (4): 1399–1440.

Hoffman, Matthew D, and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15 (1): 1593–1623.

Hoffman, Matthew D., and Andrew Gelman. 2014. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.* 15 (1): 1593–1623. http://dl.acm.org/citation.cfm?id=2627435.2638586.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Hou, Kewei, Chen Xue, and Lu Zhang. 2017. "Replicating Anomalies." National Bureau of Economic Research.

Huberman, Gur. 1982. "A Simple Approach to Arbitrage Pricing Theory." *Journal of Economic Theory* 28 (1): 183–91.

Kan, Raymond, and Guofu Zhou. 2007. "Optimal Portfolio Choice with Parameter Uncertainty." *Journal of Financial and Quantitative Analysis* 42 (3): 621–56.

Klein, Roger W, and Vijay S Bawa. 1976. "The Effect of Estimation Risk on Optimal Portfolio Choice." *Journal of Financial Economics* 3 (3): 215–31.

Lewandowski, Daniel, Dorota Kurowicka, and Harry Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100 (9): 1989–2001.

Lo, Andrew W, and A Craig MacKinlay. 1990. "Data-Snooping Biases in Tests of Financial Asset Pricing Models." *The Review of Financial Studies* 3 (3): 431–67.

MacKay, David J. C. 2003. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press.

McLean, R David, and Jeffrey Pontiff. 2016. "Does Academic Research Destroy Stock Return Predictability?" *The Journal of Finance* 71 (1): 5–32.

Meager, Rachael. 2019. "Understanding the Average Impact of Microcredit Expansions: A Bayesian Hierarchical Analysis of Seven Randomized Experiments." *American Economic Journal: Applied Economics* 11 (1): 57–91.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *Journal of Chemical Physics* 21 (June): 1087–92. https://doi.org/10.1063/1.1699114.

Neal, Radford M. 1996. "Sampling from Multimodal Distributions Using Tempered Transitions." *Statistics and Computing* 6 (4): 353–66. https://doi.org/10.1007/BF00143556.

Rubin, Donald B. 1981. "Estimation in Parallel Randomized Experiments." *Journal of Educational Statistics* 6 (4): 377–401.

Vasicek, Oldrich A. 1973. "A Note on Using Cross-Sectional Information in Bayesian Estimation of Security Betas." *The Journal of Finance* 28 (5): 1233–9.

Vehtari, Aki, Jonah Gabry, Yuling Yao, and Andrew Gelman. 2019. "Loo: Efficient Leave-One-Out Cross-Validation and Waic for Bayesian Models." https://CRAN.R-project.org/package=loo.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and Waic." *Statistics and Computing* 27 (5): 1413–32.