

RETURN FORECASTS AND OPTIMAL PORTFOLIO CONSTRUCTION: A QUANTILE REGRESSION APPROACH

LINGJIE MA AND LARRY POHLMAN

ABSTRACT. In finance there is growing interest in quantile regression with the particular focus on value at risk and copula models. In this paper, we first present a general interpretation of quantile regression in the context of financial markets. We then explore the full distributional impact of factors on returns of securities, and find that factor effects vary substantially across quantiles of returns. Utilizing distributional information from quantile regression models, we propose two general methods for return forecasting and portfolio construction. We show that under mild conditions these new methods provide more accurate forecasts and potentially higher value-added portfolios than the classical conditional mean method.

Keywords: return forecast, quantile regression, portfolio construction

Version: June 8, 2007. We thank Roger Koenker for his very helpful comments and encouragement. We also thank the comments from John Minahan, two anonymous referees of the *European Journal of Finance*, and participants in the following seminars: Forecasting Financial Markets in Marcelles, France; PanAgora Research Seminar; Northfield Annual Summer Research Conference, Financial Management Association Annual Conference at Chicago. Contact information: Lingjie Ma, Northern Trust Global Investment, lm107@ntrs.com; Larry Pohlman, Wellington Management Company, lfpohlman@wellington.com.

1. INTRODUCTION

There is intensive research on return forecasts for securities, and most of it has used the conditional mean estimation strategy. The classical least squares methods and maximum likelihood estimators provide attractive methods of estimation for Gaussian linear equation models with additive errors. However, these methods offer only a conditional mean view of the relationship, implicitly imposing quite restrictive location-shift assumptions on the way that covariates are allowed to influence the conditional distributions of the response variables. Quantile regression methods seek to broaden this view, offering a more complete characterization of the stochastic relationship among variables and providing more robust, and consequently more efficient, estimates in some non-Gaussian settings.

Since the seminal paper of Koenker and Bassett (1978), quantile regression has gradually become a complimentary approach for the traditional conditional mean estimation methods. However, in the financial markets, quantile regression has not been employed until quite recently. Most of applications of quantile regression in finance have been focused on conditional value at risk models. While the current literature of quantile regression in finance has focused on risk (VaR model) or ex-post analysis of return, in this paper we focus on return forecasting and also discusses portfolio construction. We use the quantile regression to explore the distributional effects of factors on the response of equity models. We introduce two general strategies to illustrate how the distributional information can be utilized to construct an optimal portfolio and show that the portfolio resulting from the distributional estimation method outperforms the portfolio constructed based on the classical conditional mean estimation methods.

Section 2 presents a brief review of the literature on equity models and the applications of quantile regression methods. Section 3 describes the details of quantile regression method and gives some examples. Section 4 introduces a general interpretation of quantile regression in equity models as well as the advantages and challenges compared to the traditional methods. Section 5 proposes two approaches for the construction of optimal portfolio using the distributional information. Section 6 concludes.

2. REVIEW

One question remains central in financial markets, “Is it possible to forecast the return of securities?” Understanding of the financial markets has been attributed to numerous authors including (but not limited to) Sharpe (1964, CAPM), Fama and French (1992, three-factor model), Lakonishok et al. (1992, behavior finance). A typical quantitative approach for the return forecasting is to build a linear or non linear model based on a combination of valuation, technical, and expectational factors and apply various conditional mean techniques such as least squares methods for the estimation.

Indeed, applying the standard conditional mean estimation strategy to a multi-factor model across a diverse range of stocks is a popular approach to forecasting returns. However this approach forces returns of all securities to have the homogeneous mean-reversion behavior. This mean treatment effect (MTE) view of the factor effect is valid under the extremely strong condition that the *average* marginal effect of a factor does not vary across the size of factor and returns. However for models of equity returns this condition does not in general hold. For example this condition implies that a leverage ratio of total liabilities to assets will affect the relative high return and low return securities with exactly the same direction and same magnitude. But we know that for a successful firm (one with higher returns), the leverage ratio will play a marginal role in affecting its returns while for a struggling firm (one with lower returns) due to say financial distress, the leverage ratio will have a much large effect on its operation and hence returns. The MTE view is not able to capture such characteristics of factor effects.¹ Realizing that the factor effects are not constant across securities, some studies, such as Barnes and Hughes (2002), have pointed out that an estimate of a distributional effect would be preferred. Quantile regression is a natural tool to accomplish such a task.

A primary example of the growing interest for quantile regression in finance is in the context of risk management, as witnessed by the literature on Value at Risk. Since VaR is simply a particular quantile of future portfolio values conditional on current information, the quantile regression is a natural tool to tackle such a problem. Engle and Manganelli (1999) were among the first to consider the quantile regression for the VaR model. They construct a conditional autoregressive value at risk model (CAViAR), and employ quantile regression for the estimation. To evaluate the goodness of fit for the estimated results, they also propose a test based on the idea in Chernozhukov (1999). Their applications to real data suggested that the tails follow different behavior from the middle of the distribution, which contradicts the assumption behind GARCH and risk metrics since such approaches implicitly assume that the tails follow the same process as the rest of the returns. Following Engle and Manganelli (1999), Giacomini and Komunjer (2002) propose a testing procedure for the analysis of competing conditional quantile regression forecasts proposed in the literature for the VaR model.

Recently, Chen and Chen (2003) conducted an empirical study to compare performances of Nikkei 225 VaR calculations with the quantile regression approach to those with the conventional variance-covariance approach. They find that VaR calculations with quantile regression outperform those with the variance-covariance approach; furthermore, the advantage of the quantile regression is more significant for longer holding periods.

¹By specifying certain function forms for a factor, the MTE view may reveal some variations of factor effects across different values of factors.

In addition to the application in VaR models, there are studies employing quantile regression in other areas of financial market. Barnes and Hughes (2002) employ quantile regression to test whether the conditional CAPM holds at points of the return distribution other than the mean. Their empirical results provide new support for Merton's (1987) prediction that size and returns are positively related and is in contrast to many of the empirical findings reported in the literature. Realizing that it would be a mistake to think a single measure should describe a portfolio's style, Bassett and Chen (2001) employ quantile regression to characterize the distributional view of portfolio management styles. They find that large positive and negative impacts in the tails can cancel each other which leads to barely significant results from the conditional mean estimates.

While the use of quantile regression in various financial fields has enriched the understanding of financial market, one great potential would be to apply this new tool to the quantitative investing practice. To the best of our knowledge, there has been no study employing quantile regression for the purpose of return forecasting and portfolio construction. As an initial effort on this topic, this paper explores some basic methods which might be used to link the quantile regression to portfolio construction. We illustrate the quantile regression with a set of variables commonly used in equity models to provide a general view on how quantile regression could be used for a better return forecasting and portfolio construction.

3. QUANTILE REGRESSION: A BRIEF INTRODUCTION

Why do we need quantile regression? In most cases, the effects of factors on response are not constant, but rather vary across the responses. Suppose for example that we would like to know: how does the debt-ratio factor affect a stock's return at high and low value? By using the conditional mean estimation method, we would come to the conclusion that in spite of different return levels, the debt ratio will affect the returns of all securities exactly the *same* way. However, if there is heterogeneity in the effects it will not be captured by the conditional mean estimation method.

How do we capture such heterogeneous effects? To answer this question, consider for example the following simple linear model:

$$(3.1) \quad y_{i,t} = x_{i,t}^\top \beta + u_{i,t},$$

where $y_{i,t}$ is the return and $x_{i,t}$ is the k -dimension vector of factors of the stock i at time t , $i = 1, 2, \dots, N$, $t = 1, 2, \dots, T$. Assume x is independent of u . For simplicity, consider only the cross-section version of (3.1), i.e., t is fixed at only one period. For notation convenience, we will hereafter drop the reference to the index t unless it is needed for clarification.

For the pure location shift model (3.1), under the conditions that u_i is i.i.d. and Gaussian, the traditional conditional mean estimation strategy would yield the consistent and efficient estimates for the effect of x on y . However, such restrictive

conditions of i.i.d. and Gaussian errors rarely hold in real market data. It is usually true that the effect of x on y would not be constant across y . For example,

$$(3.2) \quad y_i = x_i^\top \beta + \phi(x_i, \gamma) \epsilon_i;$$

where $\phi(\cdot)$ is usually unknown. Note that now $u_i = \phi(x_i, \gamma) \epsilon_i$ is not i.i.d. even if ϵ_i is assumed to be i.i.d. How does the location-scale shift model, (3.2), capture the distributional effects of x on y ?

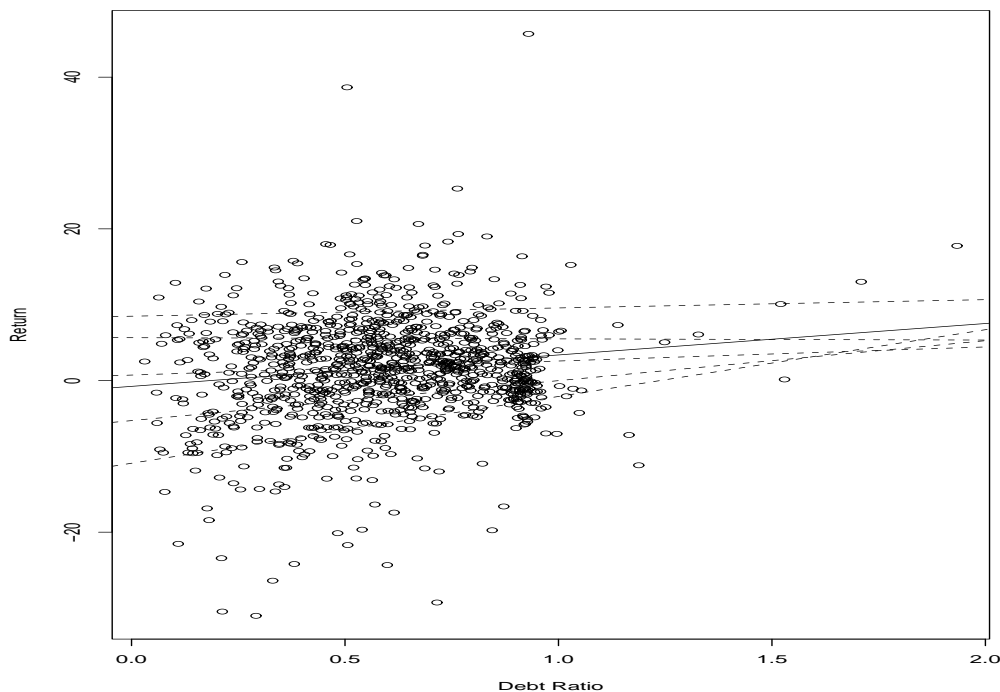


FIGURE 3.1. Heterogeneous Effects of Debt Ratio on Return. The data is from the May, 2004 of Valueline. The solid line is the OLS result while the dashed lines are the quantile regression results at $\tau = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. The different slopes of quantile regression lines indicate that the debt ratio has quite different impacts on the returns at different levels.

The quantile regression approach provides the answer. For the purpose of illustration, consider the simplest single factor version of (3.2),

$$(3.3) \quad y_i = \beta_0 + \beta_1 x_i + (x_i \gamma) \epsilon_i.$$

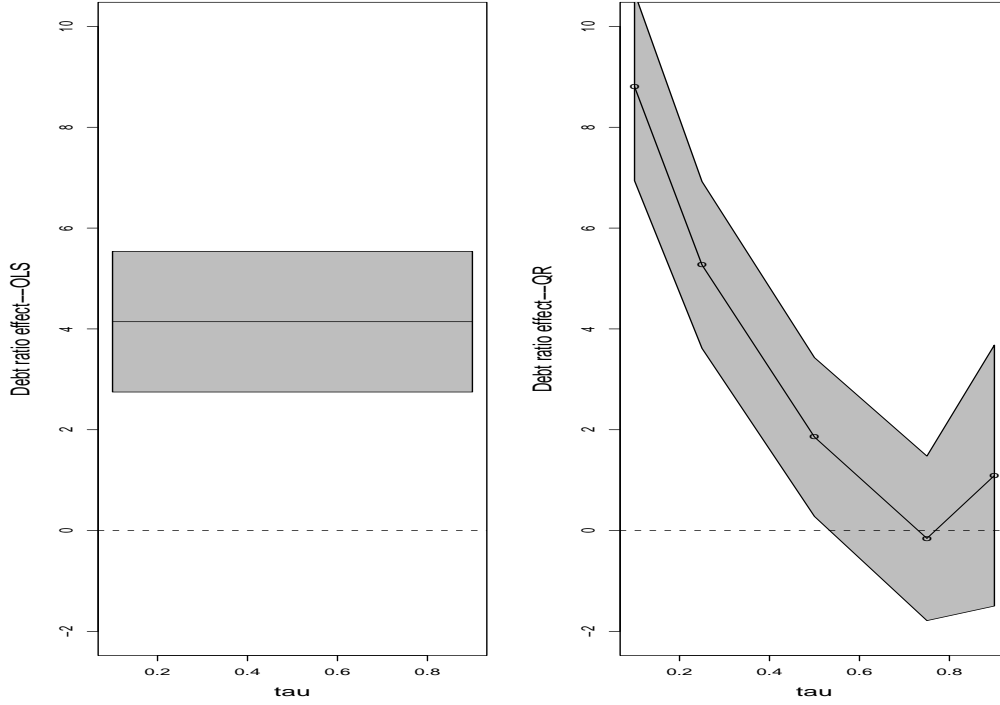


FIGURE 3.2. Heterogeneous Effects of Debt Ratio on Return. The OLS and QR results are described in the left and right figure, respectively. The five points are the slope estimates at $\tau = \{0.1, 0.25, 0.5, 0.75, 0.9\}$ and the gray area indicates the 90% confidence band.

For (3.3), assuming ϵ_i is independently distributed with the distribution function F_i , we have

$$\begin{aligned}
 (3.4) \quad F_{y_i}^{-1}(\tau|x_i) &= \beta_0 + \beta_1 x_i + (x_i \gamma) F_i^{-1}(\tau) \\
 &= \beta_0 + (\beta_1 + \gamma F_i^{-1}(\tau)) x_i
 \end{aligned}$$

where $\tau \in (0, 1)$. Correspondingly, the conditional quantile regression model is

$$(3.5) \quad Q_{y_i}(\tau|x_i) = \beta_0(\tau) + \beta_1(\tau)x_i,$$

where $\beta_0(\tau) = \beta_0$ and $\beta_1(\tau) = \beta_1 + \gamma F_i^{-1}(\tau)$. Thus, with τ varying within the range $(0, 1)$, model (3.5) presents a distributional view of the effect of debt ratio on return.

The above point is illustrated vividly in Figure 3.1–3.2, where the effects of debt ratio on returns are explored by using both the OLS and QR methods. Figure 3.1 is a scatter plot with slopes obtained from OLS and QR at $\tau = \{0.1, 0.25, 0.5, 0.75, 0.9\}$. There is dramatic variation of the effects of debt ratio across quantiles of conditional return distribution. The magnitude of the heterogeneity can be seen more clearly from Figure 3.2. While both OLS and QR results indicate that increase of debt ratio will

benefit return, the QR tells a more detailed story on how debt ratio affects returns. The left plot of OLS result implies that the debt ratio has a uniformly significant marginal effect of 4.1 on the return. The right plot of QR results suggests that such effects are not at all constant: the effect is as high as 9 for the companies with returns in left tail and becomes insignificant as the conditional quantile of return passes the median. Although this is just from a one-factor model, the intuition is clear that the effects of debt ratio are heterogeneous across returns.

The parameter $\beta(\tau) = (\beta_0(\tau), \beta_1(\tau))^\top$ in model (3.5) can be estimated by solving the following conditional quantile objective function:

$$(3.6) \quad \min_{b \in \mathcal{B}} \sum_{i=1}^n \rho_\tau(y_i - \tilde{x}_i^\top b),$$

where $\tilde{x}_i = (1, x_i)$ and the check function $\rho(\cdot)$ is defined as $\rho_\tau(e) = (\tau - I(e \leq 0))e$ and $e = y_i - \tilde{x}_i^\top b$.²

Noted that QR is not the same as commonly used sorts on independent variables. Instead QR can be thought of as sorting on the dependent variable (for example the returns) conditional on factors and then estimating the parameters for different parts of the return distribution. QR is not the same as running regressions of subsets of the data: all QR estimates depend upon all of the observations, which is shown clearly in (3.6).

A natural question is under what conditions QR will be better than OLS in terms of full characterization of the stochastic relationship between variables. If there is heterogeneity then QR will provide a more complete view of the relationship between variables through the effects of independent variables across quantiles of the response distribution. If the conditional return distribution is not Gaussian but fat tailed, then QR estimates will be more robust and efficient than the conditional mean estimates.

4. A GENERAL INTERPRETATION OF QR IN EQUITY MODEL

In this section, we first present a very general interpretation of QR in equity models and then illustrate the approach in detail through a popular multi-factor model. We discuss the advantages of QR over classic methods and the challenges of using the distribution information for return forecasting and portfolio construction.

4.1. A General Interpretation. Consider the semiparametric model with a general form,

$$(4.1) \quad y = G(x, \beta, u),$$

²The check function is the short notation for the objective function:

$$\sum_{e_i \geq 0} \tau e_i + \sum_{e_i < 0} (\tau - 1)e_i.$$

where y is the return of security, x is the k -dimension vector of factors, u is the error term, which is assumed to be i.i.d. with distribution function F . Note that x may include the lagged response variables. We assume that x is independent of u . The function $G(\cdot)$ is rather flexible to allow for either a linear or nonlinear specification.

Under mild conditions³, we can write the conditional quantile functions,

$$(4.2) \quad \begin{aligned} Q_y(\tau|x) &= G(x, \beta, F_u^{-1}(\tau)) \\ &= G(x, \beta(\tau)). \end{aligned}$$

How do we interpret (4.2) in financial markets? One approach is to regard (4.2) as a natural model for value at risk, which is defined as the value that a portfolio will lose with a given probability (τ). However (4.2) explains more than just VaR. Note that model (4.2) could be interpreted as the value at risk, or the return depending on the emphasis. Without any context, model (4.2) gives the probability of return at τ th percentile conditional on x is $\beta(\tau)$. In financial markets, $\beta(\tau)$ shows how the information contained in x affect the returns at different points of distribution. Thus, model (4.2) presents a full view of the effects or forecasting power of the selected factors on returns. Shifting the focus from τ to $\beta(\tau)$ shifts the application of QR from risk to return.

4.2. An Empirical Example. We illustrate the points of the above general interpretation through a simple linear multi-factor model. We use monthly observations from Valueline database for US equity market for the period January 1990 to June 2004.⁴ We focus on the large cap market, which consists of stocks in the Russell1000, S&P500 and S&P400 indices for about 1,100 securities each month. We use the one month forward return as the dependent variable and the securities value, technical and expected information as the independent variables. Note that the variables used in the paper are very common factors studied in the finance literature and used by practitioners.⁵ The value regressors include the book to price (BTOP), earnings to price (ETOP), debt ratio (Dratio), retained earnings to assets (REOA), liabilities to

³The detailed conditions for (4.2) are specified in more detail in the following:

- A.1:** The conditional distribution functions $F_{y_i}(y_i|x_i)$ is absolutely continuous with continuous densities f_i that is uniformly bounded away from 0 and ∞ at the points $\xi_i = Q_y(\tau|x_i)$ for $i = 1, \dots, n$.
- A.2:** The function $G(\cdot)$, is assumed strictly monotonic in u , and differentiable with respect to x .

⁴The financial ratios are calculated based on quarterly data where available and annual otherwise.

⁵For example, Fama and French (1992) examined value (book value to market value) and size effects, Altman (1968) studies a zscore for the distress effect, Jegadeesh (1990) documented the short term reversal effect.

price (LTOP) and zscore (Zscore).⁶ For the technical factors the variables include a reversal factor measured by 3-month lag return (*Ret_L3*) and market cap (*VL_CAP*). For the expected information factor, we control for earnings per share for the next year from IBES (*IBES_EPS*).

A brief statistical summary of the variables is reported in Table A.1. in the Appendix. The quantiles, means and standard deviations for the variables illustrate a number of interesting characteristics of the sample. First is that the variation of all variables (except for ETOP and Dratio) are very high, indicating that identification will not be a problem. Even for the factors of BTOP and Dratio, the range is very high and both factors have symmetric distributions. For the technical factors there is a considerable variation of firm sizes measured by market cap, and the mean is 3 times of the median suggesting a large skewness toward the right tail. At the expectational factor level the most interesting fact is that all the forecasts of one-year forward IBES_EPS are positive. It appears the analysts are all extremely optimistic about future earnings.

We employ the following simple linear quantile regression model,

$$(4.3) \quad Q_y(\tau|x) = x^\top \beta(\tau),$$

where x includes an intercept, the value factors including BTOP, ETOP, Dratio, REOA, LTOP, and Zscore, the technical factors *Ret_L3* and *VL_CAP* and the expectational factors from IBES, *IBES_EPS*.

Model (4.3) is estimated at $\tau = \{0.1, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8, 0.9\}$. The results of parameter estimates are depicted in Figure 4.1 with 90% confidence band. OLS results are also plotted as bold lines in Figure 4.1. An overall impression suggests that QR reveals significant and interesting results that would have been hidden in the traditional model.

Consider first the effects of value factors on the forward return. For the BTOP factor the effect is not significant at the left tail of the return distribution but as τ increases the effect increases and is about 4 at $\tau = 0.8$. Note that the OLS result is close to the median which is barely significant. For the debt ratio the pattern of effects is just the opposite of that of BTOP: the increase of Dratio has positive effects on the companies having lower returns and then as the return increases the effects decrease and become insignificant after the median. Although we expect that ETOP will behave similarly to BTOP we find that ETOP is insignificant except at the left tail. For the factor of REOA the effects are insignificant at the median but very significant at tails meaning that the increase of REOA has a very positive effect on

⁶Note that earnings are measured by operating income before depreciation and amortization. Liabilities in LTOP are the total reported liabilities. The debt ratio is calculated as total liabilities divided by total assets. Zscore is the linear combination of several factors indicators to measure the firm's financial strength (Altman (1968)). In this paper, Zscore is calculated as follows: $Zscore = 3.3ROA + 0.999SOA + 0.6EOD + 1.2WKOA$, where *ROA* is return on total assets, *SOA* is sales on assets, *EOD* is equity on debt, *WKOA* is working capital on assets.

lower returns but negative effects on high returns. For LTOP we find that the increase of LTOP will further reduce the value of the lower-return companies but benefit the higher-return ones. The Zscore reflects the financial strength which is expected to have positive effects on the returns. The increase of Zscore is effective only in right tails (it is barely significant at the left tail). The higher the Zscore the larger the impact on the high-return stocks. For OLS, the Zscore is not significant at all.

At the technical level we find that the reversal factor does work as a reversal but only for the low return stocks: for high return stocks the previous return does not matter. Note that OLS is the same as the zero line. We find that size does matter but in a very heterogeneous way: the increase of size has a positive effect on lower return stocks but a negative effect on the high return stocks.

Finally, for the effects of analysts forecasts, IBES_EPS, we find that the one year forecast of the earnings-per-share has negative effect on the stocks with lower returns but no effect on the stocks with high returns. This finding suggests that betting against the analysts might not be a bad strategy.

Note that for most factors the results from OLS are not significant. The comparison of QR results with OLS results suggests that one should interpret findings of insignificant mean effects with considerable caution since it appears that those results arise from averaging significant benefits from reductions (increase) in factor values for high return stocks and significant benefits from increase (reductions) in factor values for low return stocks.

4.3. Advantages and Challenges. It is clear that with the presence of heterogeneity there are advantages of quantile regression over the traditional ones such as least squares. First of all, instead of the point estimate for the conditional mean, we have the whole distribution. With τ varying in the range of $(0, 1)$ we potentially have different $\beta(\tau)$. Therefore we would know not only the expected average return, but the whole distribution of the return in the next period given the information known at the current moment. Secondly, the conditional mean result could be derived from the conditional quantile effect. In particular if the distribution of effects are not too skewed, the conditional mean effect would be close to the median. Therefore, for the same sample, quantile regression reveals more information than classical methods. However since we are estimating more parameters we must be more cautious about our inference. Also the more information, the more challenging the decision making.

For our return forecasting and final portfolio construction purpose, the main concern is how the quantile regression method could be used to yield more reasonable forecasts and hence the optimal portfolio. For the classical approach, we have only one set of point estimates, and then one forecast for each stock for the asset return of future periods. For example, for the conditional mean approach, we have only $\hat{E}(y_{i,t+1}|x_{i,t})$ for each stock i . But now given multiple sets of forecasting returns, say, for example, at $\tau = \{0.1, 0.25, 0.5, 0.75, 0.9\}$, the forecasts of return will be $\{\hat{Q}_{y_i}(0.1|x_i), \hat{Q}_{y_i}(0.25|x_i), \hat{Q}_{y_i}(0.5|x_i), \hat{Q}_{y_i}(0.75|x_i), \hat{Q}_{y_i}(0.9|x_i)\}$. How can we take

advantage of such distribution information to have more accurate forecasts and more value-added portfolio? This is a challenging task. To the best of our knowledge for quantitative analysis of equity models, this is the first study to address and investigate such an issue. We explore the answers in the next section.

5. RETURN FORECAST AND PORTFOLIO CONSTRUCTION USING DISTRIBUTIONAL INFORMATION

We propose two methods to forecast returns and construct corresponding optimal portfolios using the distributional information from the quantile regression estimation strategy: quantile regression alpha distribution (QRAD) and quantile regression portfolio distribution (QRPD). Before going into details of QRAD and QRPD, we introduce some propositions to measure and compare the goodness of fit for different forecasts.

Given that the security returns are usually not Gaussian distributed but instead have very fat tails, the sum of absolute values is a better measure of accuracy than the more usual sum of squared differences. Also, the absolute value is robust compared to the quadratic one. Under the special case of symmetric distribution of the returns, the two measures will be identical. It should be noted that quantile regression properties of the full characterization of stochastic relationship between variables are not loss function dependent.⁷

Proposition 1. *Suppose R^m is the median return and R^e is the expected return of a set of returns R_i , $i = 1, 2, \dots, N$, then*

$$\sum_{i=1}^N |R^m - R_i| \leq \sum_{i=1}^N |R^e - R_i|.$$

Proof: The proof follows immediately from the definition of median and expected return. ■

Now, consider extending the unconditional median (expected) return to the conditional ones,

$$(5.1) \quad R_{i,t+1} = G(x_{it}, \beta, u_{i,t+1})$$

and let $\hat{R}_i^m(0.5|x)$ and $\hat{R}_i^e(\cdot|x)$ denote the forecasted returns from the conditional median and mean, say, OLS, estimate, respectively, the above proposition still holds.

Proposition 2. *Let $\hat{R}_i^m(0.5|x)$ be the median return and $\hat{R}_i^e(\cdot|x)$ be the expected return from (5.1), then*

$$\sum_{i=1}^N |\hat{R}_i^m - R_i| \leq \sum_{i=1}^N |\hat{R}_i^e - R_i|.$$

⁷This has been kindly pointed out by a referee.

Proof: The proof is straightforward and it comes directly from the minimization of the following objective function:

$$\hat{\beta}(0.5) = \operatorname{argmin}_{b \in B} \sum_{i=1}^N \rho_{\tau}(R_{i,t+1} - x_{i,t}^{\top} b) = \operatorname{argmin}_{b \in B} \sum_{i=1}^N |R_{i,t+1} - x_{i,t}^{\top} b|,$$

$$\hat{R}^m(0.5|x) = x^{\top} \hat{\beta}(0.5).$$

■

So far, for (5.1), only the median estimates are used. What if we have the distributional estimates? Is there a way that the combination would be better than the median?

Consider again model (5.1). For simplicity, let $\tau = (0.1, 0.5, 0.9)$, then the corresponding forecasting returns will be⁸

$$\hat{R}_{0,1} = x^{\top} \hat{\beta}(0.1), \quad R_{0,5} = x^{\top} \hat{\beta}(0.5), \quad R_{0,9} = x^{\top} \hat{\beta}(0.9).$$

However, for a stock i , we do not know the specific quantile of the next period return so which $\hat{\beta}(\tau)$ should be used and how accurate would be the result?

5.1. The QRAD Method. We introduce two sub approaches for the QRAD method, namely, QRAD location and QRAD probability. We will discuss each of them in detail in the following.

5.1.1. QRAD Location. The QRAD location method assumes that stocks remain in the same quantiles from period to period. The method is based upon the assumption that for the most part the *rank* of returns of stocks does not change dramatically. Thus by the QRAD-Location method, we would have the forecast for blocks of stocks that are in the same quantile. The final forecasts for all stocks in the sample would be the combination of forecasts for these blocks.

For example, suppose we choose $\tau = \{0.1, 0.5, 0.9\}$ and calculate the corresponding $\hat{\beta}(0.1), \hat{\beta}(0.5), \hat{\beta}(0.9)$. Then at time t if the conditional quantile for $R_{i,t}$ belongs to $(0, 0.1]$, then we would use $\hat{\beta}(0.1)$ to get the next period forecast $\hat{R}_{i,t+1} = x^{\top} \hat{\beta}(0.1)$. Applying the same rule, we would obtain the forecast returns for all stocks in the sample,

$$(5.2) \quad \hat{R}_{i,t+1} = \begin{cases} x_{i,t+1}^{\top} \hat{\beta}(0.9) & \text{if } R_{i,t} \geq x_{i,t}^{\top} \hat{\beta}(0.9) \\ x_{i,t+1}^{\top} \hat{\beta}(0.1) & \text{if } R_{i,t} \leq x_{i,t}^{\top} \hat{\beta}(0.1) \\ x_{i,t+1}^{\top} \hat{\beta}(0.5) & \text{otherwise} \end{cases}.$$

The goodness of fit can be decomposed as follows,

$$\sum_{i=1}^N |\hat{R}_{i,t+1} - R_{i,t+1}| = \sum_A |\hat{R}_{i,t+1} - R_{i,t+1}| + \sum_B |\hat{R}_{i,t+1} - R_{i,t+1}| + \sum_C |\hat{R}_{i,t+1} - R_{i,t+1}|,$$

⁸Note that we drop the subscript for time for the simple notation.

where A , B and C denotes the area of $(0, 0.1]$, $(0.1, 0.9)$ and $[0.9, 1)$, respectively, for the quantile location of $R_{i,t}$ (See Figure 5.1). The relative accuracy of the forecasts based on QRAD location method is stated in the following proposition.

Proposition 3. *Let $\hat{R}^l(\tau|x)$ be the composite quantile return from the QRAD location method, also let $\hat{R}^m(0.5|x)$ and $\hat{R}^e(\cdot|x)$ be the median and mean return from (5.1), respectively, then*

$$\sum_{i=1}^N |\hat{R}_{i,t+1}^l - R_{i,t+1}| \leq \sum_{i=1}^N |\hat{R}_{i,t+1}^m - R_{i,t+1}| \leq \sum_{i=1}^N |\hat{R}_{i,t+1}^e - R_{i,t+1}|,$$

where the letter l , m and e denotes location, median and mean, respectively.

Proof: We need the proof only for the first inequality since the second inequality is established by proposition 2. Let $\Delta_1 \equiv \sum_{i=1}^N |\hat{R}_{i,t}^l - R_{i,t}|$ and $\Delta_2 \equiv \sum_{i=1}^N |\hat{R}_{i,t}^m - R_{i,t}|$. The decomposition by location yields

$$\begin{aligned} \Delta_1 &= \sum_A |\hat{R}_{i,t+1}^{0.1} - R_{i,t+1}| + \sum_B |\hat{R}_{i,t+1}^m - R_{i,t+1}| + \sum_C |\hat{R}_{i,t+1}^{0.9} - R_{i,t+1}|, \\ \Delta_2 &= \sum_A |\hat{R}_{i,t+1}^m - R_{i,t+1}| + \sum_B |\hat{R}_{i,t+1}^m - R_{i,t+1}| + \sum_C |\hat{R}_{i,t+1}^m - R_{i,t+1}|. \end{aligned}$$

Then, $\Delta_1 \leq \Delta_2$ is equivalent to

$$\sum_A |\hat{R}_{i,t+1}^{0.1} - R_{i,t+1}| + \sum_C |\hat{R}_{i,t+1}^{0.9} - R_{i,t+1}| \leq \sum_A |\hat{R}_{i,t+1}^m - R_{i,t+1}| + \sum_C |\hat{R}_{i,t+1}^m - R_{i,t+1}|.$$

By the condition that for stocks with $R_{i,t} \in A$,

$$Prob.(R_{i,t+1} \in A) \geq Prob.(R_{i,t+1} \in B) \text{ and } Prob.(R_{i,t+1} \in A) \geq Prob.(R_{i,t+1} \in C),$$

we have

$$\sum_A |\hat{R}_{i,t+1}^{0.1} - R_{i,t+1}| \leq \sum_A |\hat{R}_{i,t+1}^m - R_{i,t+1}|.$$

The same inequality holds for C . The final result follows immediately from the combination. \blacksquare

The results of Proposition 3 can be extended to situations with more than three areas. However, it should be noted that there is the trade off between the number of divisions, the accuracy of forecast and the likelihood that a stock will remain in the same quantile.

5.1.2. *QRAD Probability.* To overcome the disadvantage of the QRAD location method where forecasts depend heavily on previous conditional location, we propose an alternative sub approach, which is to assign probabilities to the forecasts,

$$\hat{R}_{i,t+1}^p = p_1 \hat{R}_{1,i,t+1} + p_2 \hat{R}_{2,i,t+1} + \dots + p_k \hat{R}_{k,i,t+1},$$

where p_k is the probability of the occurrence of $\hat{R}_{k,i,t+1}$. For example, for $\tau = \{0.1, 0.5, 0.9\}$, we would have that

$$(p_1, p_2, p_3) = (0.1, 0.8, 0.1),$$

and,

$$(\hat{R}_{1,i,t+1}, \hat{R}_{2,i,t+1}, \hat{R}_{3,i,t+1}) = x_{i,t}^\top (\hat{\beta}(0.1), \hat{\beta}(0.5), \hat{\beta}(0.9)).$$

Clearly, the above formulation is the familiar expected value. However, it should be noted that the expected forecasted return is not identical to the forecast of expected return in the general case, that is, $E\hat{R} \neq \widehat{ER}$. Under the conditions of convexity, $E\hat{R} \leq \widehat{ER}$.

Again, it can be shown that,

$$\sum_{i=1}^N |\hat{R}_{i,t+1}^p - R_{i,t+1}| \leq \sum_{i=1}^N |\hat{R}_{i,t+1}^m - R_{i,t+1}| \leq \sum_{i=1}^N |\hat{R}_{i,t+1}^e - R_{i,t+1}|.$$

Under mild conditions, both QRAD-location and QRAD-probability yield better goodness of fit than the traditional methods. What is the relationship between QRAD-Location and QRAD-Probability? We find that $\hat{R}_{i,t+1}^p = E\hat{R}_{i,t+1}^l$. Note that from (5.2), under the conditions in Proposition 3, by the definition of quantile regression, the probability chart (Figure 5.2) follows. In another word, the two methods are asymptotically the same. Thus we have used more information to construct the forecast and these forecast results are more accurate than the conditional mean forecast which does not differentiate the factor effects for different level of returns.

5.2. The QRPD Method. The QRPD method differs from the QRAD method by using the distributional information at the optimization stage instead of making use of the distributional information at the forecasting stage.

Corresponding to the forecasted returns R_k at each different conditional quantile τ_k , the optimal portfolio could be constructed. Let $W_\tau = (w_{1,\tau}, \dots, w_{N,\tau})^\top$ be the optimal weights of the portfolio resulted from using the τ th quantile regression forecasted returns, then with $\tau \in (0, 1)$, we have an empirical distribution of the portfolio at time t . Using the same strategy as for the QRAD-probability method, the final QRPD portfolio at time t is constructed as follows

$$W = p_1 W_{\tau_1} + \dots + p_k W_{\tau_k},$$

where p_k is the probability of occurrence of W_{τ_k} . For example, as $\tau = \{0.1, 0.5, 0.9\}$, we would have three sets of weights, $W_{0.1}$, $W_{0.5}$ and $W_{0.9}$, corresponding to the three sets of forecasted returns. Thus, the weights of the proposed portfolio will be

$$W = 0.1W_{0.1} + 0.8W_{0.5} + 0.1W_{0.9}.$$

With this new portfolio construction methodology, there are two questions we need to answer. First, what is the relationship between portfolios constructed from QRAD forecasts and QRPD? Second, does the portfolio constructed from the QRPD method

outperform those from the median or mean forecasts? That is, does the inequality, $W_q^\top R \geq W_m^\top R$, hold?

To explore the answer to the first question, consider the following utility function,

$$(5.3) \quad \max_W W^\top R - \lambda W^\top \Omega W,$$

where Ω is the covariance matrix of R and λ is the risk acceptance parameter. We want to derive the sufficient conditions that the methods of QRAD-Probability and QRPD yield the same portfolio.

Suppose R_k is the forecasted returns from the τ_k quantile regression, then by QRAD-probability, we have

$$R = \sum_{i=1}^k p_i R_i$$

Furthermore, since the focus of our research is the impact of the return forecast on the portfolio we will assume that we use the same risk model

$$\Omega_i = \Omega$$

for all of the optimizations. This is not unusual since risk forecasting models are regarded separately from return forecasts.

The first order condition of (5.3) yields

$$\begin{aligned} W_{qrp} &= \frac{1}{2\lambda} \Omega^{-1} R \\ &= \frac{1}{2\lambda} \left(\sum p_i \Omega_i^{-1} R_i \right) \\ &= \sum p_i W_i \\ &= W_{gradp}. \end{aligned}$$

In the unconstrained case with a common risk model we have shown that using the quantile regression distribution information at the forecasting or portfolio construction stage will give the same result.

The second question is answered by the following proposition.

Proposition 4. *Suppose R_1 and R_2 are two sets of forecasting returns for the same set of stocks with real return R , and R_1 is better than R_2 by the goodness of fit:*

$$(5.4) \quad |R_1 - R| \leq |R_2 - R|.$$

then under conditions that $\Omega_1 = \Omega_2 = \Omega$, where Ω_j is the covariance matrix, we have that,

$$(5.5) \quad W_1^\top R_1 \geq W_2^\top R_2,$$

where W_j is derived from

$$\max W_j^\top R_j - \lambda W_j^\top \Omega_j W_j.$$

Proof: We'll prove the result by tracing back from (5.5) to (5.4). Suppose W is the weights for R , then we have

$$(5.6) \quad W^\top R \geq W_1^\top R \geq W_2^\top R.$$

By the condition that $W_j = \frac{1}{2\lambda}\Omega_j^{-1}R_j$, (5.6) is equivalent to

$$\begin{aligned} W^\top R &\geq \frac{1}{2\lambda}R_1^\top\Omega^{-1}R \geq \frac{1}{2\lambda}R_2^\top\Omega^{-1}R \\ \iff W^\top R &\geq W^\top R_1 \geq W^\top R_2 \\ \iff 0 &\leq W^\top R - W^\top R_1 \leq W^\top R - W^\top R_2 \\ \iff 0 &\leq W^\top(R - R_1) \leq W^\top(R - R_2) \\ \iff |W^\top R - W^\top R_1| &\leq |W^\top R - W^\top R_2| \end{aligned}$$

A strong sufficient condition for the last inequality is that
for the long positions ($w_i > 0$), $R_i - R_{2i} \geq R_i - R_{1i} \geq 0$,
for the short positions ($w_i < 0$), $R_i - R_{2i} \leq R_i - R_{1i} \leq 0$.
Thus (5.4) follows immediately:

$$|R_{1i} - R_i| \leq |R_{2i} - R_i|.$$

■

Remarks

(i) Condition (5.4) is the strongest measure of goodness of fit. A rather weaker condition is:

$$(5.7) \quad Prob(|R_{1i} - R_i| \leq |R_{2i} - R_i|) \geq 0.$$

(ii) However, condition (5.7) is still quite strong from the practical point of view. A rather general condition is

$$\sum |R_1 - R| \leq \sum |R_2 - R|,$$

which, although is not sufficient to yield the optimal portfolio, implies a higher chance. A better forecast does not necessarily yield a better portfolio.

Suppose we relax the restriction on risk models or add constraints such as position limits. The answer is then no longer analytic. However, in general, if the objective function is convex then by Jensen's inequality we expected that the following inequality of portfolio value holds: $V_{QRAD-Probability} \geq V_{QRPD} \geq V_{Median} \geq V_{Expected}$.

6. CONCLUSION

Equity return forecasting and portfolio construction continue to be two significant issues in both the academic and industrial worlds and statistical estimation plays central role. However, most of the analytical approaches are based on the conditional mean method, which ignores the heterogeneity of the effects of factors on returns.

In this paper we focus on the heterogeneity issue from the response side and introduce quantile regression as a natural statistical tool to tackle such an issue. We present a general interpretation of the quantile regression results for equity models by expanding the interpretation to not only conditional risk but the conditional return as well. Regarding quantile regression as a general alternative approach to the classical conditional mean method, we then focus on return forecasting and portfolio construction taking advantage of the distribution information from quantile regression. The main challenge is how to utilize the distribution information to construct more accurate forecast and better performing portfolios given that the quantile of future return is unknown. To accomplish such tasks, we propose two methods, QRAD and QRPD, where the former utilizes the distributional information at the forecasting stage while the latter at the portfolio construction stage.

By using the goodness of fit measure, we show that results from both QRAD and QRPD outperform the results from traditional methods. Further empirical research is required to quantify the gain from applying quantile regression to the two problems of forecasting and portfolio construction.

REFERENCES

- [1] Altman, E.: "Financial Ratios, Discriminate Analysis and the Prediction of Corporate Bankruptcy", *Journal of Finance*, September, 1968.
- [2] Amemiya, T.: "Two Stage Least Absolute Deviations Estimators", *Econometrica*, 50, 689–711, 1982.
- [3] Barnes, M and Hughes, A.: "A Quantile Regression Analysis of the Cross Section of Stock Market Returns", working paper, Federal Reserve Bank of Boston, 2002.
- [4] Bassett, G and Chen, H.: "Portfolio Style: Return-Based Attribution Using Quantile Regression", *Empirical Economics*, Springer-Verlag, pp. 1405–1441, 2001.
- [5] Blundell, R and Powell, J.: "Endogeneity in Nonparametric and Semiparametric Regression Models", in *Advances in Economics and Econometrics: Theory and Applications*, Eighth World Congress, Cambridge University Press, 2003.
- [6] Chen, M and Chen, J.: "Application of Quantile Regression to estimation of value at Risk", working paper, 2003.
- [7] Chesher, A.: "Identification in Nonseparable Models", *Econometrica*, vol. 71, No. 5, pp. 1405–1441, 2003.
- [8] Engle, R. and Manganelli, S.: "CAViaR: Conditional Autoregressive Value at Risk by Regression Quantiles", *Journal of Business and Economic Statistics*, 2004.
- [9] Fama, E. and French, K.: "The Cross-section of Expected Stock Returns", *Journal of Finance*, vol. 47, pp. 427-465, 1992.
- [10] Ihaka, R. and Gentleman, R.: "R, A Language for Data Analysis and Graphics", *Journal of Graphical and Computational Statistics*, 5, 299-314, 1996.
- [11] Jegadeesh, N.: "Evidence of predictable behavior of security returns", *Journal of Finance*, 45, 881-898, 1990.
- [12] Koenker, R. and Bassett, G.: "Regression Quantiles", *Econometrica*, 46, 33–50, 1978.
- [13] Koenker, R. and Park, B.: "An Interior Point Algorithm for Nonlinear Quantile Regression", *Journal of Econometrics*, 71, 265-285, 1996.
- [14] Koenker, R. "Quantreg: A Quantile Regression Package for R," <http://cran.r-project.org>, 1998.
- [15] Koenker, R.: "Quantile Regression", *Cambridge University Press*, 2005.
- [16] Kordas, G.: "Smoothed Binary Quantile Regression", *Journal of Applied Econometrics*, 2004.
- [17] Kordas, G.: "Credit Scoring Using Binary Quantile Regression", in *Statistical Data Analysis Based on the L1-Norm and Related Methods*, Yalolah Dodge (editor), 2002, Birkhauser.
- [18] Konno, H. and Yamazaki, H.: "Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock Market", *Management Science*, Vol. 37, No. 5, 519-531, 1991.
- [19] Lakonishok, J., Shleifer, A. and Vishny, R.: "The Impact of Institutional Trading on Stock-prices", *Journal of Financial Economics*, vol. 32, August, pp. 2343, 1992.
- [20] Ma, L. and Koenker, R.: "Quantile Regression Methods for Recursive Structural Equation Models", *Journal of Econometrics*, 2006.
- [21] Sharpe, W.: "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk", *Journal of Finance*, vol. 19 (3), pp. 425-442, 1964.
- [22] Zhao, Q.: "Asymptotically Efficient Median Regression in the Presence of Heteroscedasticity of Unknown Form", *Econometric Theory*, 17, 765-84, 2001.

7. TABLE FOR SUMMARY OF FACTOR STATISTICS

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Std.
Return	-98.00	-3.06	1.62	2.28	6.78	212.30	10.33
Valuation							
BTOP	-35.54	0.25	0.41	0.47	0.60	19.30	0.61
Dratio	0.03	0.42	0.60	0.60	0.78	2.83	0.25
ETOP	-7.76	0.05	0.10	0.13	0.17	11.44	0.26
REOA	-31.29	0.04	0.17	0.12	0.38	1.47	1.04
LTOP	0.01	0.25	0.60	1.76	1.62	377.80	5.59
Zscore	-39.75	1.30	2.82	4.15	4.91	127.60	5.81
Technical							
<i>Ret_L3</i>	-76.08	-3.38	5.57	7.30	15.21	423.80	20.26
<i>VL_CAP</i>	14.23	1,604	2,946	9,751	7,362	339,600	25,193
Expectational							
<i>IBES_EPS</i>	1.00	7.00	12.00	13.41	18.00	45.00	7.90

TABLE 7.1. Summary for response and factor statistics. The statistics describes the distribution of return and factor values: minimum, maximum, mean standard deviation, first quartile, median and third quartile. The factors listed are employed in a multifactor model in Section 4.2.

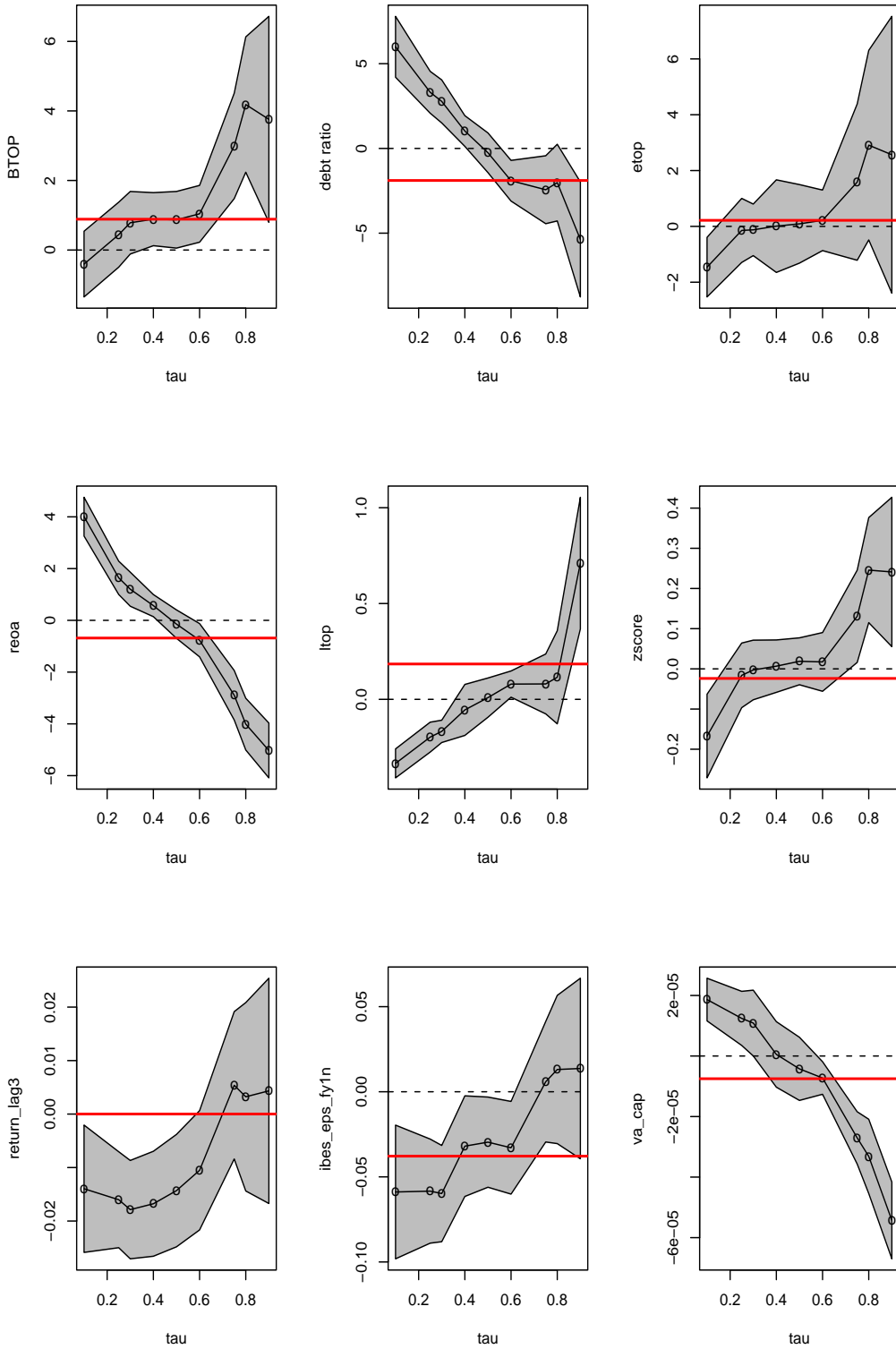


FIGURE 4.1. Quantile regression results at $\tau = \{0.1, 0.25, 0.3, 0.4, 0.5, 0.6, 0.75, 0.8, 0.9\}$. The gray area is the 90% confidence band. The plots are based on the estimates from the multi-factor model (4.3) in Section 4. The horizontal solid line is the result from OLS. The dashed zero line is for the visual convenience of coefficient significance.

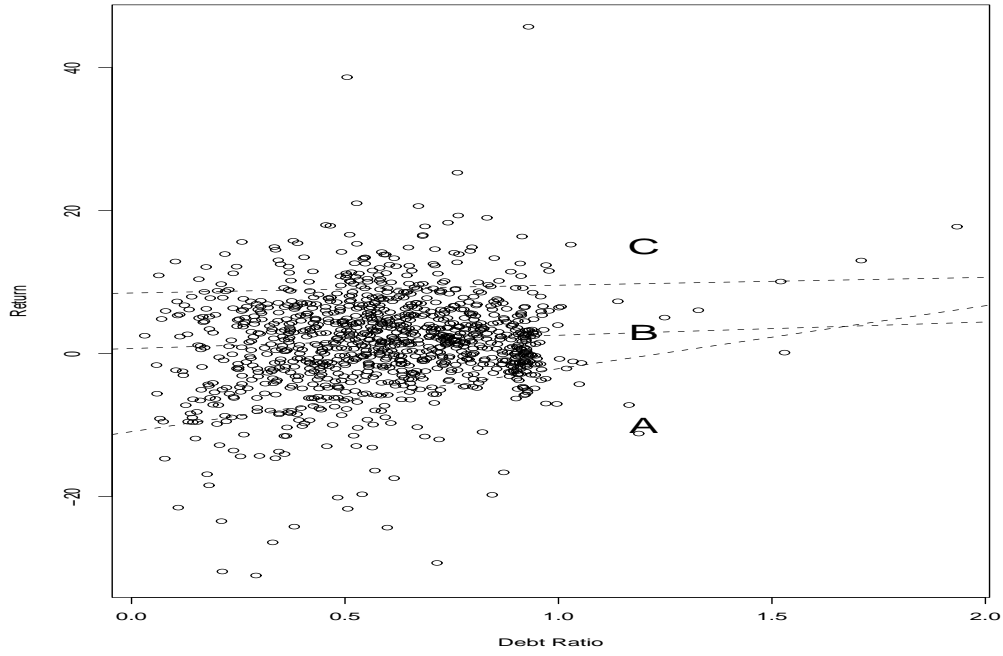


FIGURE 5.1. QRAD-Location. The dashed lines are the quantile regression results at $\tau = \{0.1, 0.5, 0.9\}$. The letter A, B and C indicates the area below the line with $\tau = 0.1$, between the lines with $\tau = 0.1$ and $\tau = 0.9$, and above the line with $\tau = 0.9$, respectively.

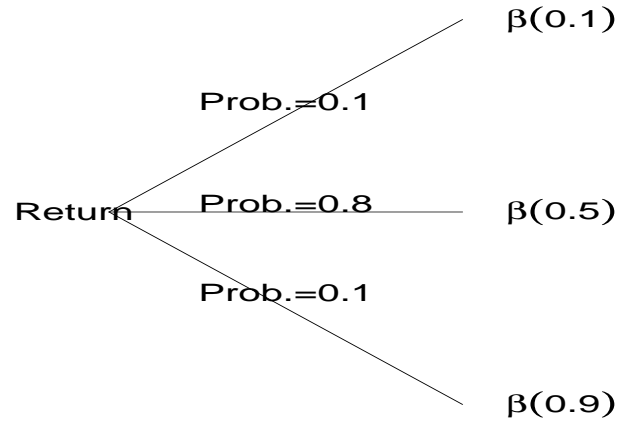


FIGURE 5.2. Probability of QRAD-Location. By the definition of quantile regression, $\beta(\tau)$ is the effect of a factor on the τ^{th} quantile of return distribution. The probability of a stock's return below (above) 10% (90%) return distribution conditioning on factors is 10%. Therefore, $\beta(0.1)$, $\beta(0.9)$ and $\beta(0.5)$ are assigned with probability of 10%, 10% and 80%, respectively, for the stock return forecasts for the QRAD-Probability method.