AIMS Mathematics

http://www.aimspress.com/journal/Math

AIMS Mathematics, 8(6): 13123–13133.
DOI:10.3934/math.2023662
Received: 18 February 2023
Revised: 16 March 2023
Accepted: 19 March 2023
Published: 03 April 2023

Research article

Linear regression estimation using intraday high frequency data

Wenhui Feng, Xingfa Zhang*, Yanshan Chen and Zefang Song

School of Economics and Statistics, Guangzhou University, Guangzhou 510006, China

* **Correspondence:** Email: xingfazhang@gzhu.edu.cn.

Abstract: Intraday high frequency data have shown important values in econometric modeling and have been extensively studied. Following this point, in this paper, we study the linear regression model for variables which have intraday high frequency data. In order to overcome the nonstationarity of the intraday data, intraday sequences are aggregated to the daily series by weighted mean. A lower bound for the trace of the asymptotic variance of model estimator is given, and a data-driven method for choosing the weight is also proposed, with the aim to obtain a smaller sum of asymptotic variance for parameter estimators. The simulation results show that the estimation accuracy of the regression coefficient can be significantly improved by using the intraday high frequency data. Empirical studies show that introducing intraday high frequency data to estimate CAPM can have a better model fitting effect.

Keywords: linear regression; intraday high frequency data; quadratic programming
Mathematics Subject Classification: 62J05, 62M10

1. Introduction

With the development of electronic technology, intraday high frequency data become easily available. Such data are valuable in statistical modeling and financial risk assement [1–4]. By utilizing high frequency data, potential risks can be identified more efficiently and accurately due to an increased level of detail in the analysis of the markets. However, due to the nonstationarity and periodicity of the intraday high frequency data, generally it is not appropriate to directly introduce these data into a stationary model [5]. A possible way is aggregating the intraday high frequency data to a daily stationary quantity or constructing a low frequency stationary proxy [6, 7]. To illustrate this idea, we plot three return series related to CSI (China Shanghai-Shenzhen) 300 index in Figure 1. We collected the returns of the CSI 300 index from 01 Sep 2017 to 12 July 2019, including 466 daily observations. There are 240 observations each day based on the intrady sampling frequency of 1 min. Subplot (a) in the Figure 1 is the time series plot of the intraday sequence on the seventh day of the data set, which shows an obvious time trend and nonstationarity; subplot (b) is the time series plot of the seventh intraday observation for the first 240 days of the data set; subplot (c) is the time series plot of the mean of intraday sequences for the first 240 days of the data set. It is seen that the series in subplots (b) and (c) tend to be stationary sequences, which implies the fact that although the intraday sequence can be nonstationary while its aggregation (weighted mean) might be stationary.

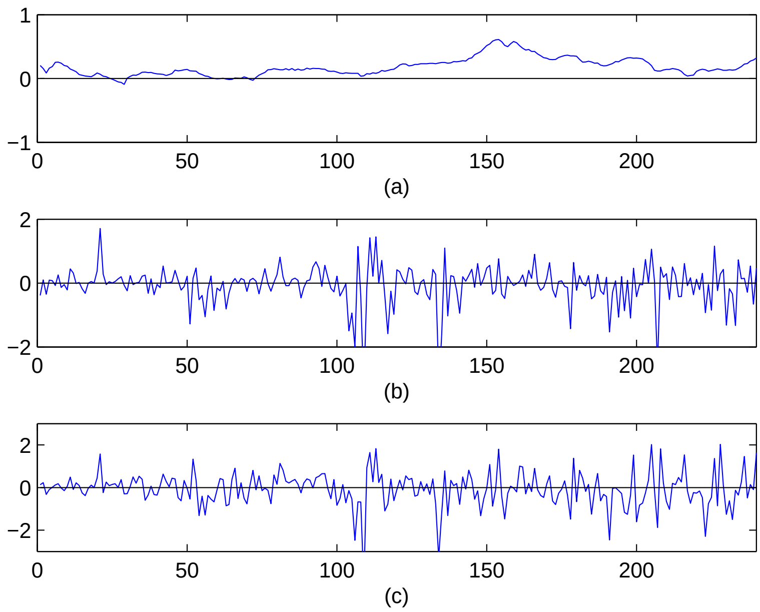

Figure 1. (a) is the time series plot of the intraday sequence on the seventh day of the data set; (b) is the time series plot of the seventh intraday observation for the first 240 days of the data set; (c) is the time series plot of the mean of intraday sequences for the first 240 days of the data set.

Linear regression model has been extensively applied in the area of daily financial time series analysis, such as ARMA model, linear pricing model, factor model and other linear forecasting models, see [8–18]. Let y_t be the observation of dependent variable and $X_t = (1, x_{t1}, \cdots, x_{tp})^\tau$ be the the observation of independent variable vector at day t. Then the classic linear regression model has the form

$$y_t = X_t^\tau \beta + \varepsilon_t, \tag{1.1}$$

where $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^\tau$ is the regression coefficient vector and ε_t is the error term. To avoid the spurious regression problem, $\{y_t, X_t\}$ are normally assumed to be stationary [19].

The main goal of this paper is to study the following linear regression model with intraday high frequency data

$$y_t(u_i) = X_t^\tau(u_i)\beta + \varepsilon_t(u_i), \tag{1.2}$$

where $y_t(u_i), X_t^\tau(u_i), \varepsilon_t(u_i)$ are the observations at time u_i on the t-th day, $1 \le i \le k$, u_i is the scaled time with $0 \le u_i \le 1$. When $u_i = 1$, the series become the daily sequence, namely, $(y_t(1), X_t^\tau(1), \varepsilon_t(1)) = (y_t, X_t^\tau, \varepsilon_t)$. $\varepsilon_t(u)$ is assumed to be an independent and identically distributed errors process with zero mean and finite variance for each fixed u. For model (1.2), one can directly use the daily data (y_t, X_t^τ) to estimate the coefficient without introducing the intraday data information. However the information is not efficiently used in this occasion. Alternatively, one can aggregate the intraday high frequency data to a daily quantity and then obtain a more precise parameter estimator for the model.

Different from the well known mixed data sampling (MIDAS) regression model of Ghysels et al. [20], the dependent and independent variables in model (1.2) have the same sampling frequency which makes the regression coefficients keep unchanged after the high frequency data are aggregated by a weighted mean form. Such a property enables us to estimate low frequency regression

$$y_t = X_t^\tau \beta + \varepsilon_t, \text{namely}, y_t(1) = X_t^\tau(1)\beta + \varepsilon_t(1)$$

by taking the intraday high frequency data into account.

The contributions of this paper are as follows. First, this paper proposes a linear regression model which aggregates the intraday high-frequency data to a daily quantity. Second, a lower bound for the trace of the asymptotic variance of model estimator is given. Third, we propose a simple data-driven method for choosing the weight for aggregation of the high frequency data, with the aim to obtain a smaller sum of asymptotic variance for parameter estimators. Different from the existent methods, the weight is not restricted to certain parametric form and can be obtained by simple restricted quadratic programming.

The rest of the paper is organized as follows. Section 2 introduces the model and estimation. Section 3 investigates the estimation performance based on simulation studies. An empirical study is provided in Section 4. We conclude the paper in Section 5.

2. Model and estimation

2.1. A lower bound for least squared estimator

For model (1.1), define

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} X_1^\tau \\ X_2^\tau \\ \vdots \\ X_n^\tau \end{pmatrix}, \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{pmatrix}.$$

The least squared estimator for β is given by

$$\hat{\beta} = (X^\tau X)^{-1} X^\tau Y \tag{2.1}$$

and under regularity conditions, the following asymptotic normality holds:

$$\hat{\beta} \sim N(\beta, (X^\tau X)^{-1}\sigma^2), \tag{2.2}$$

where σ^2 is the variance of ε_t in (1.1) and it is also equivalent to $\text{var}(y_t|X_t)$. In practice, we always hope the asymptotic variance of each parameter will not be too large. Equivalently, we hope the trace of the asymptotic variance matrix $tr[(X^\tau X)^{-1}\sigma^2]$ is small. The following proposition gives an approximate lower bound for $tr[(X^\tau X)^{-1}\sigma^2]$ based on the samples $\{y_t, X_t\}_{t=1}^n$.

Proposition 1. *Suppose $\{y_t, X_t\}$ are stationary processes and all the eigenvalues of $X^\tau X$ are positive, for fixed sample size n, then a lower bound for $tr[var(\hat{\beta})]$ is given by*

$$\frac{var(y_t|X_t)}{E(X_t^\tau X_t)}\frac{(p+1)^2}{n}(1+o_p(1)).$$

Proof. Before the statement of detailed proof, we first list two properties of matrix trace.

P1 Suppose A is a symmetric $m\times m$ matrix and all the the eigenvalues of A are positive, then $tr(A^{-1}) \geq \frac{m^2}{tr(A)}$.

P2 Suppose A is a $s \times m$ matrix and B is a $m \times s$ matrix, then $tr(AB) = tr(BA)$.

According to P1,

$$tr[(X^\tau X)^{-1}] \geq \frac{(p+1)^2}{tr(X^\tau X)}. \tag{2.3}$$

By ergodicity theorem for stationary time series,

$$X^\tau X = n\frac{1}{n}\sum_{t=1}^{n} X_t X_t^\tau = n[E(X_t X_t^\tau) + o_p(1)]$$

and

$$\frac{(p+1)^2}{tr(X^\tau X)} = \frac{(p+1)^2}{n}\frac{1+o_p(1)}{tr[E(X_t X_t^\tau)]}. \tag{2.4}$$

Further, according to P2,

$$tr[E(X_t X_t^\tau)] = E[tr(X_t X_t^\tau)] = E[tr(X_t^\tau X_t)] = E[X_t^\tau X_t]. \tag{2.5}$$

Recall $tr[\mathrm{var}(\hat{\beta})] = tr[(X^\tau X)^{-1}\sigma^2] = tr[(X^\tau X)^{-1}\mathrm{var}(y_t|X_t)]$. Then the result of Proposition 1 is proved based on (2.3)–(2.5).

2.2. Aggregation of intraday high frequency data

Denote $\{y_t(u_i), x_{t1}(u_i), \cdots, x_{tp}(u_i), \varepsilon_t(u_i)\}$ to be observations at time u_i on the t-th day, $1 \leq i \leq k$, namely there are k intraday observations for each variable. For demonstration, we rewrite (1.2) as followed:

$$y_t(u_i) = \beta_0 + \beta_1 x_{t1}(u_i) + \beta_2 x_{t2}(u_i) + \cdots + \beta_p x_{tp}(u_i) + \varepsilon_t(u_i). \tag{2.6}$$

Let $y_t^* = \sum_{i=1}^{k} y_t(u_i)w_i$, $x_{tq}^* = \sum_{i=1}^{k} x_{tq}(u_i)w_i (q = 1, 2, \cdots, p)$, $\varepsilon_t^* = \sum_{i=1}^{k} \varepsilon_t(u_i)w_i$, $\sum_{i=1}^{k} w_i = 1, w_i \geq 0$. From (2.6),

$$y_t^* = \beta_0 + \beta_1 x_{t1}^* + \beta_2 x_{t2}^* + \cdots + \beta_p x_{tp}^* + \varepsilon_t^*. \tag{2.7}$$

It is easy to see that ε_t^* is still i.i.d sequence with zero mean and finite variance based on the assumption on $\varepsilon_t(u_i)$ and the daily sequences $y_t^*, x_{tq}^*(q = 1, 2, \cdots, p)$ are supposed to be stationary after aggregation. Consequently, $\beta_0,\beta_1, \cdots ,\beta_p$ can be estimated based on $y_t^*, x_{tq}^*(q = 1, 2, \cdots, p)$ and we denote the corresponding estimator as $\tilde{\beta}$. It is hoped that $\tilde{\beta}$ would be more precise than $\hat{\beta}$ which only uses low frequency information. To construct proper $y_t^*, x_{tq}^*(q = 1, 2, \cdots, p)$ is equivalent to find a proper weight vector $w = (w_1, \cdots, w_{p+1})^\tau$. Next, we give a method to choose the weight w.

Define

$$Y_t(u) = \begin{pmatrix} y_t(u_1) \\ y_t(u_2) \\ \cdots \\ y_t(u_k) \end{pmatrix}, X_t(u) = \begin{pmatrix} x_{t0}(u) \\ x_{t1}(u) \\ \cdots \\ x_{tp}(u) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{t1}(u_1) & x_{t1}(u_2) & \cdots & x_{t1}(u_k) \\ \cdots & & & \\ x_{tp}(u_1) & x_{tp}(u_2) & \cdots & x_{tp}(u_k) \end{pmatrix}.$$

Then

$$X_t^* = X_t(u)w = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{t1}(u_1) & x_{t1}(u_2) & \cdots & x_{t1}(u_k) \\ \cdots & & & \\ x_{tp}(u_1) & x_{tp}(u_2) & \cdots & x_{tp}(u_k) \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \cdots \\ w_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{k} x_{t0}(u_i)w_i \\ \sum_{i=1}^{k} x_{t1}(u_i)w_i \\ \cdots \\ \sum_{i=1}^{k} x_{tp}(u_i)w_i \end{pmatrix}, y_t^* = w^\tau Y_t(u) = \sum_{i=1}^{k} y_t(u_i)w_i.$$

According to Proposition 1, $\tilde{\beta}$ based on $y_t^*, x_{tq}^*(q = 1, 2, \cdots, p)$ has the property that

$$tr[\text{var}(\tilde{\beta})] \geq \frac{\text{var}(y_t^*|X_t^*)}{E(X_t^{*\tau}X_t^*)} \frac{(p+1)^2}{n}(1 + o_p(1)).$$

Intuitively, if the above right bound is smaller, then we can expect to obtain a less $tr[\text{var}(\tilde{\beta})]$. Such an intuition gives a way to choose the weight: finding a w which can get the smallest value for $\text{var}(y_t^*|X_t^*)/E(X_t^{*\tau}X_t^*)$. Note that when X_t^* is deterministic, we have $\text{var}(y_t^*|X_t^*) = \text{var}(y_t^*)$. On the other hand, from

$$\text{var}(y_t^*) = E[\text{var}(y_t^*|X_t^*)] + \text{var}[E(y_t^*|X_t^*)],$$

smaller $\text{var}(y_t^*)$ will bring smaller $\text{var}(y_t^*|X_t^*)$. Consequently, the rule to choose the weight can be transformed to that: finding a w which can get the smallest value for $\text{var}(y_t^*)/E(X_t^{*\tau}X_t^*)$. Further,

$$\frac{\text{var}(y_t^*)}{E(X_t^{*\tau}X_t^*)} = \frac{w^\tau \text{var}(Y_t(u))w}{w^\tau E(X_t^\tau(u)X_t(u))w}. \tag{2.8}$$

Let $\Sigma_{yu} = \text{var}(Y_t(u)), \Sigma_{xu} = E(X_t^\tau(u)X_t(u))$. From the above, if the quantity

$$\frac{w^\tau \Sigma_{yu} w}{w^\tau \Sigma_{xu} w}$$

is small, then $\tilde{\beta}$ is supposed to be more precise. In practice, Σ_{yu} and Σ_{xu} can be respectively estimated by their corresponding sample variance or sample mean, namely,

$$\hat{\Sigma}_{yu} = \frac{1}{n}\sum_{t=1}^{n}(Y_t(u) - \bar{Y})(Y_t(u) - \bar{Y})^\tau, \hat{\Sigma}_{xu} = \frac{1}{n}\sum_{t=1}^{n} X_t^\tau(u)X_t(u),$$

where $\bar{Y}$ is the sample mean vector of $Y_t(u)$. Hence the weight vector w is chosen as the minimizer of the following objective function

$$\arg\min_{w \in R^{p+1}} \frac{w^\tau \hat{\Sigma}_{yu} w}{w^\tau \hat{\Sigma}_{xu} w} \tag{2.9}$$

such that $\sum_{i=1}^{k} w_i = 1, w_i \geq 0$.

Theoretically, without restriction $w_i \geq 0$, the solution of w in (2.9) is the eigenvector of the smallest eigenvalue for matrix $\hat{\Sigma}_{xu}^{-1}\hat{\Sigma}_{yu}$, denoted as $\hat{w}$. Hence $\hat{w}$ is also the solution of

$$\arg\min_{w \in R^{p+1}} w^{\tau}\hat{\Sigma}_{xu}^{-1}\hat{\Sigma}_{yu}w. \tag{2.10}$$

From the above, the solution in (2.9) can be approximated by the solution of the quadratic programming in (2.10) with restrictions: $\sum_{i=1}^{k} w_i = 1, w_i \geq 0$. More detail about the quadratic programming with restrictions can be referred to Huyer and Neumaier [21]. The simulation studies in the following section show that such an approximation works well.

3. Simulation

In this section, we assess the finite-sample performance of the proposed estimator $\tilde{\beta}$. The sample was simulated from the model below

$$y_t(u_i) = 0.1 + 0.3x_{t1}(u_i) + 0.4x_{t2}(u_i) + \varepsilon_t(u_i), \tag{3.1}$$

$1 \leq i \leq 20, u_i = i/20$. Following Visser [7] , $x_{t1}(u_i)$ and $x_{t2}(u_i)$ were independently simulated from the following process $\xi_t(u)$ with different parameter setting:

$$\begin{aligned} d\gamma_t(u) &= -\delta(\gamma_t(u) - \mu)du + \sigma_\gamma dB_t^{(2)}(u), \\ d\xi_t(u) &= e^{\gamma_t(u)}dB_t^{(1)}(u), u \in [0, 1]. \end{aligned} \tag{3.2}$$

The Brownian motions $B_t^{(1)}$ and $B_t^{(2)}$ were uncorrelated, $\xi_t(0) = 0$, and $\gamma(0)$ was sampled from $N(\mu, \sigma_r^2)$. We divided the unit time interval $[0, 1]$ into 20 small intervals, set $\delta = 1/2$, $\sigma_\gamma = 1/4$, $\mu = -1/16$ for $x_{t1}(u)$ and $\delta = 1/3$, $\sigma_\gamma = 1/5$, $\mu = -1/14$ for $x_{t2}(u)$. $\varepsilon_t(u_i) \sim i.i.dN(0, 0.64)$, and then $y_t(u_i)$ can be obtained based on (3.1). When $u_i = 1$, we also get the daily sample $\{y_t(1), x_{t1}(1), x_{t2}(1), \varepsilon_t(1)\}$, namely $\{y_t, x_{t1}, x_{t2}, \varepsilon_t\}$, such that

$$y_t = 0.1 + 0.3x_{t1} + 0.4x_{t2} + \varepsilon_t. \tag{3.3}$$

Let $\tilde{\beta}$ and $\hat{\beta}$ be the estimator from (3.1) and (3.3) respectively. Here $\tilde{\beta}$ introduces the intraday high frequency information, as discussed in Section 2, while $\hat{\beta}$ only uses the daily sequence. Hence, $\tilde{\beta}$ is expected to be more precise than $\hat{\beta}$. The sample sizes of n = 50, 100 and 150 are considered, and the replication time is 1000. Table 1 reports the sample bias and the sample standard deviation of $\tilde{\beta}$ and $\hat{\beta}$, denoted as BS1, BS2, SD1 and SD2, respectively. From the table, we can receive several observations as follows. The biases of $\tilde{\beta}$ are smaller than those of $\hat{\beta}$, and both become smaller when the sample size n increases. This implies that both estimators are asymptotically unbiased. The sample standard deviations of $\tilde{\beta}$ are also significantly smaller than those of $\hat{\beta}$, and both become smaller when the sample size n increases. This implies that $\tilde{\beta}$ performs better than $\hat{\beta}$ does in our simulations.

Let $S_\sigma \equiv \text{var}(y_t^*)/E(X_t^{*\tau}X_t^*)$. From (2.8), for smaller S_σ, a less $tr[\text{var}(\tilde{\beta})]$ is expected. To justify this expectation, for each replication, S_σ for $\tilde{\beta}$ and $\hat{\beta}$ are respectively estimated by sample variance and sample mean for each sample size. Figure 2 shows the box plots of the S_σ series. It can be seen that the median S_σ values of $\tilde{\beta}$ are smaller than those of $\hat{\beta}$, and both become smaller when the sample size n increases. Such a result is consistent with Table 1, justifying the intuition: the smaller S_σ is, the less $tr[\mathrm{var}(\tilde{\beta})]$ would be. According to the simulation results, it is shown that introducing the intraday high frequency data can significantly improve the estimation of the regression coefficient.

Table 1. Bias and standard deviation of the estimator.

sample size	BS1	BS2	SD1	SD2
		Result for β_0		
$n = 50$	−0.0027	−0.0037	0.0791	0.1239
$n = 100$	−0.0026	0.0003	0.0488	0.0789
$n = 150$	0.0001	0.0035	0.0365	0.0700
		Result for β_1		
$n = 50$	−0.0073	−0.0034	0.0066	0.0132
$n = 100$	−0.0048	−0.0038	0.0039	0.0086
$n = 150$	−0.0038	−0.0034	0.0030	0.0071
		Result for β_2		
$n = 50$	−0.0091	−0.0074	0.0052	0.0124
$n = 100$	−0.0064	−0.0070	0.0034	0.0087
$n = 150$	−0.0051	−0.0072	0.0025	0.0071

†Number of replications=1000.

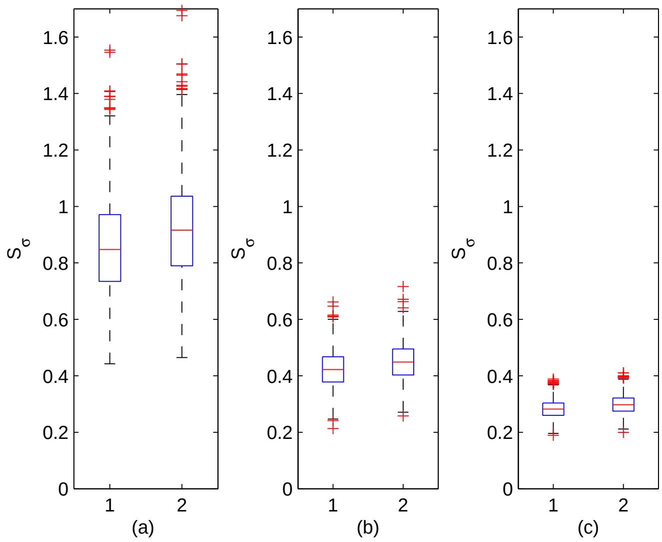

Figure 2. Subplots $(a), (b)$ and (c) are the box plots of S_σ series for $\tilde{\beta}$ and $\hat{\beta}$ (from left to right in each subplot) under sample size n=50, 100 and 150 respectively.

4. Empirical study

In this section, the proposed method is applied to study the relationship between single stock and stock index based on the classic Capital Asset Pricing Model (CAPM), see Latunde et al. [22]. Denote $P_t(u)$ as the t-th intraday price sequence. We calculate the intraday log-return as

$$Y_t(u) = 100[\log P_t(u) - \log P_{t-1}(1)], u \in [0, 1]. \tag{4.1}$$

According to the CAPM, we choose the Shanghai Composite Index as the market portfolio and randomly choose JCHX Mining Management (stock code: 603979) as the single asset, from China Shanghai Stock Exchange. Let $rm_t(u)$ and $r_t(u)$ be the intraday high frequency return series for Shanghai Composite Index and JCHX Mining stock respectively, which can be computed based on (4.1). And rm_t and r_t, namely $rm_t(1)$ and $r_t(1)$ are the daily return series. Classic CAPM implies the following relationship between rm_t and r_t:

$$r_t = \beta_0 + \beta_1\, rm_t + \varepsilon_t, \tag{4.2}$$

and β_1 is the famous beta coefficient used in the CAPM, and it describes the relationship between systematic risk and expected return for assets (usually stocks). The beta coefficient can be used to help investors understand whether a stock moves in the same direction as the rest of the market. It also provides insights into how risky a stock is relative to the rest of the market. Consequently, it makes sense to get a more precise estimation for the beta coefficient by using extra information. Following this motivation, we introduce the intraday high frequency data in model (4.2):

$$r_t(u) = \beta_0 + \beta_1\, rm_t(u) + \varepsilon_t(u), \tag{4.3}$$

Note that models (4.2) and (4.3) share the same regression coefficient while (4.3) takes intraday high frequency information into account and can have a more precise estimation for the parameters, as discussed in the Sections 2 and 3.

For the considered series $rm_t(u)$ and $r_t(u)$, the data span the period from 19 Nov 2019 to 17 Jan 2020, which consist of 43 daily observations. For each day, the intraday sampling frequency: 1min, 5min, 15min, 30min and 60min are considered. We can get the estimations for models (4.2) and (4.3) by applying the method given in Section 2. Table 2 lists the results for (4.2) (in the last column) and (4.3) with different sampling frequency, where CIL and CIU denote the 95% confidence lower bound and upper bound respectively, R^2 is the R-squared coefficient of linear regression, Tr is the estimated values for $\mathrm{var}(\hat{\beta}_0) + \mathrm{var}(\hat{\beta}_1)$ in Table 2.

From Table 2, different sampling frequencies get different beta values from 0.4429 to 0.7149. According to the computed R^2 and Tr, the fitting effect is the best under the sampling frequency of 1min and the estimated beta value is 0.5341, which is smaller than the value 0.6379 estimated only by the daily information.

The above results imply that using the intraday high frequency data to estimate the CAPM can have a better model fitting effect and this is helpful for the investor to make rational decisions. And such empirical studies can be easily extended to other pricing models such as ICAPM and factor model, see [23–26].

Table 2. Estimation results for (4.2) and (4.3) with different sampling frequency.

results	1min	5min	15min	30min	60min	daily
$\hat{\beta}_0$	−0.1038	−0.1319	−0.1418	−0.1185	−0.1035	0.1553
CIL for $\hat{\beta}_0$	−0.1858	−0.2240	−0.2692	−0.2921	−0.2822	−0.1026
CIU for $\hat{\beta}_0$	−0.0218	−0.0398	−0.0143	0.0551	0.0752	0.4132
$\hat{\beta}_1$	0.5314	0.5355	0.4429	0.6260	0.7149	0.6379
CIL for $\hat{\beta}_1$	0.2429	0.2273	0.0464	0.2070	0.3093	0.2519
CIU for $\hat{\beta}_1$	0.8199	0.8436	0.8394	1.0451	1.1206	1.0239
R^2	0.2524	0.2310	0.1104	0.1817	0.2360	0.2136
Tr	0.0288	0.0322	0.0467	0.0602	0.0616	0.0656

5. Conclusions

It is valuable to introduce high frequency data into low frequency standard models. These data can provide insights into trends, patterns, and correlations that may not be visible with lower frequency data. Additionally, they can help identify anomalies or outliers that may indicate risk. Analyzing high frequency data makes it possible to detect subtle changes or shifts.

The linear regression model for variables which have intraday high frequency data is studied in this paper. A method is given to estimate the model based on the idea of time series aggregation. Simulation results show the proposed approach performs well. Empirical studies imply that our model can have many potential applications in linear forecasting models.

Our research findings will provide insights for studying other linear or nonlinear time series models, such as threshold autoregression models. The method can be applied to different pricing and factor models in our future study, and it is expected to perform better.

Acknowledgments

This work is partially supported by Guangdong Basic and Applied Basic Research Foundation 2022A1515010046 (Xingfa Zhang) and Funding by Science and Technology Projects in Guangzhou SL2022A03J00654 (Xingfa Zhang), 202201010276 (Zefang Song).

Conflict of interest

The authors declare no conflict of interest.

References

1. L. Li, X. Zhang, Y. Li, C. Deng, Daily GARCH model estimation using high frequency data, (Chinese), *J. Guangxi Norm. Univ. Nat. Sci.*, **39** (2021), 68–78. http://dx.doi.org/10.16088/j.issn.1001-6600.2020091601
2. L. Li, X. Zhang, C. Deng, Y. Li, Quasi maximum exponential likelihood estimation of GARCH model based on high frequency data, (Chinese), *Acta. Math. Appl. Sin.*, **45** (2022), 652–664.

3. X. Hui, B. Sun, I. SenGupta, Y. Zhou, H. Jiang, Stochastic volatility modeling of high-frequency CSI 300 index and dynamic jump prediction driven by machine learning, *Electron. Res. Arch.*, **31** (2023), 1365–1386. http://dx.doi.org/10.3934/era.2023070

4. Z. Li, B. Mo, H. Nie, Time and frequency dynamic connectedness between cryptocurrencies and financial assets in China, *Int. Rev. Econ. Financ.*, **86** (2023), 46–57. http://dx.doi.org/10.1016/j.iref.2023.01.015

5. T. G. Andersen, T. Bollerslev, Intraday periodicity and volatility persistence in financial markets, *J. Empir. Financ.*, **4** (1997), 115–158. http://dx.doi.org/10.1016/S0927-5398(97)00004-2

6. F. C. Drost, T. E. Nijman, Temporal aggregation of GARCH processes, *Econometrica*, **61** (1993), 909–927. http://dx.doi.org/10.2307/2951767

7. M. P. Visser, GARCH parameter estimation using high-frequency data, *J. Financ. Economet*, **9** (2011), 162–197. http://dx.doi.org/10.1093/jjfinec/nbq017

8. Z. Li, L. Chen, H. Dong, What are bitcoin market reactions to its-related events?, *Int. Rev. Econ. Financ.*, **73** (2021), 1–10. http://dx.doi.org/10.1016/j.iref.2020.12.020

9. J. Saleemi, COVID-19 and liquidity risk, exploring the relationship dynamics between liquidity cost and stock market returns, *National Accounting Review*, **3** (2021), 218–236. http://dx.doi.org/10.3934/NAR.2021011

10. D. G. Kirikos, An evaluation of quantitative easing effectiveness based on out-of-sample forecasts, *National Accounting Review*, **4** (2022), 378–389. http://dx.doi.org/10.3934/NAR.2022021

11. Z. Li, H. Dong, C. Floros, A. Charemis, P. Failler, Re-examining bitcoin volatility: a CAViaR-based approach, *Int. Rev. Econ. Financ.*, **58** (2022), 1320–1338. http://dx.doi.org/10.1080/1540496X.2021.1873127

12. Z. Li, C. Yang, Z. Huang, How does the fintech sector react to signals from central bank digital currencies?, *Financ. Res. Lett.*, **50** (2022), 103308. http://dx.doi.org/10.1016/j.frl.2022.103308

13. H. Siddiqi, Financial market disruption and investor awareness: the case of implied volatility skew, *Quant. Financ. Econ.*, **6** (2022), 505–517. http://dx.doi.org/10.3934/QFE.2022021

14. T. Li, J. Wen, D. Zeng, K. Liu, Has enterprise digital transformation improved the efficiency of enterprise technological innovation? A case study on Chinese listed companies, *Math. Biosci. Eng.*, **19** (2022), 12632–12654. http://dx.doi.org/10.3934/mbe.2022590

15. Y. Liu, P. Failler, Y. Ding, Enterprise financialization and technological innovation: Mechanism and heterogeneity, *PLoS ONE*, **17** (2022), e0275461. http://dx.doi.org/10.1371/journal.pone.0275461

16. Y. Liu, P. Failler, Z. Liu, Impact of environmental regulations on energy efficiency: a case study of China's air pollution prevention and control action plan, *Sustainability*, **14** (2022), 3168. http://dx.doi.org/10.3390/su14063168

17. Y. Liu, L. Chen, L. Lv, P. Failler, The impact of population aging on economic growth: a case study on China, *AIMS Mathematics*, **8** (2023), 10468–10485. http://dx.doi.org/10.3934/math.2023531

18. Y. Liu, J. Liu, L. Zhang, Enterprise financialization and R&D innovation: A case study of listed companies in China, *Electron. Res. Arch.*, **31** (2023), 2447–2471. http://dx.doi.org/10.3934/era.2023124

19. C. Y. Choi, H. Ling, M. Ogaki, Robust estimation for structural spurious regressions and a Hausman-type cointegration test, *J. Econometrics*, **142** (2008), 327–351. http://dx.doi.org/10.1016/j.jeconom.2007.06.003
20. E. Ghysels, V. Kvedaras, V. Zemlys-Balevic, Mixed data sampling (MIDAS) regression models, *Handbook of Statistics*, **42** (2011), 162–197. http://dx.doi.org/10.1016/bs.host.2019.01.005
21. W. Huyer, A. Neumaier, MINQ8: general definite and bound constrained indefinite quadratic programming, *Comput. Optim. Appl.*, **69** (2018), 351–381. http://dx.doi.org/10.1007/s10589-017-9949-y
22. T. Latunde, S. L. Akinola, D. D. Dare, Analysis of capital asset pricing model on Deutsche bank energy commodity, *Green. Finance*, **2** (2020), 20–34. http://dx.doi.org/10.3934/GF.2020002
23. M. Stutzer, Style investing and the ICAPM, *Quant. Financ. Econ.*, **2** (2018), 702–716. http://dx.doi.org/10.3934/QFE.2018.3.702
24. Y. Liu, Z. Li, M. Xu, The influential factors of financial cycle spillover: evidence from China, *Emerg. Mark. Financ. Tr.*, **56** (2020), 1336–1350. http://dx.doi.org/10.1080/1540496X.2019.1658076
25. F. Corradin, M. Billio, R. Casarin, Forecasting economic indicators with robust factor models, *National Accounting Review*, **4** (2022), 167–190. http://dx.doi.org/10.3934/NAR.2022010
26. S. A. Gyamerah, B. E. Owusu, E. K. Akwaa-Sekyi, Modelling the mean and volatility spillover between green bond market and renewable energy stock market, *Green Finance*, **4** (2022), 310–328. http://dx.doi.org/10.3934/GF.2022015

AIMS Press