Basic Statistical Methods for Fund Selection

Introduction

For retail investors, fund represents a viable form of investment as a replacement for direct involvement in individual securities. Nonetheless, the fund market is usually interspersed with an enormous variety of funds and fund managers, where the sheer amount and diversity of information frequently cause confusion, and sometimes cultivate blind faiths in fund performance within retail investors. To counter these, it is of interest to develop a systematic approach to evaluate the styles and performances for various fund managers, such that fund selection can be made in a rational manner.

Section 1. Basic considerations on the reliability of fund managers

Generally, the most important information we want about a fund manager concerns whether he or she demonstrates superior performance over his or her peers, and whether such superiorities are stable. Superiority of performance is usually a manifestation of the manager's experience and acumen, whereas stability of performance is usually related to his or her perception of risk, which manifests as the extent of leverage of his or her positions, as well as the ability to adjust his or her exposure in response to the market environment.

Therefore, it is not hard to see that the decision to invest in a fund, just as most other investments, cannot be done without a reasonable forecast of the overall situation of the market. It is only when such a forecast is at hand that we proceed to the formulation of suitable statistical models, which aim to describe the nature and the extent of correlation between the fund manager and the overall fund market. Since our concerns involve both superiority and stability of performance, we would naturally contemplate some regression models that account for these factors, which shall then be our main topic of discussion.

Section 2: Mathematical considerations to model fund performance

To commence, when we search for relevant information on a fund manager's historical performance, we can usually access such information as

- a list of the returns for the manager's all past funds, denoted as $\{R_i\}_{i=1}^{n}$,
- a list on the time of initiation and termination for each fund, denoted as $\{t_{i,1}\}_{i=1}^{n}$ and $\{t_{i,2}\}_{i=1}^{n}$, respectively, as well as
- a list of the returns for the overall fund market, which we denote as $\{r_i\}_{i=1}^{n}$,

where n denotes the sample size for past funds. In most cases, this sample size would be small, and a relevant model should be parsimonious and make full use of the above-mentioned information. Before we furnish the mathematical details, we shall address some additional model assumptions. In this write-up, we assume that the returns over time are compounded, so that it is easier to work with the transformed returns of the form $y_i = \ln(R_i + 1)$ and $x_i = \ln(r_i + 1)$ for fund and market performances respectively. Upon such transformations, we assume that the random components of fund performance are independent for different funds and different time periods. Also, we assume that the random components for y_i follow normal distributions, which is also a common practice in financial analysis. With these assumptions, we formulate the relevant statistical models on the mean annual performance of a fund. To do this, let $\bar{x}_i = \frac{x_i}{t_{i,2}-t_{i,1}}$ and $\bar{y}_i = \frac{y_i}{t_{i,2}-t_{i,1}}$. The linear model for mean annual performance then becomes

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \bar{\varepsilon}_i$$

and the quadratic model becomes

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + \beta_2 (\bar{x}_i - m_{\bar{x}_i})^2 + \bar{\varepsilon}_i,$$

where $m_{\bar{x}_i}$ denotes the sample mean of mean annual market performance. The purpose for the centralization is to reduce multicollinearity between the linear and quadratic terms, which stabilizes parameter estimates. Both models can be rewritten in matrices with

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y}$ denotes the vector of mean annual performances, $\mathbf{X}$ denotes the data matrix which may or may not include the centralized quadratic term, $\boldsymbol{\beta}$ denotes the parameter estimate vector, and $\boldsymbol{\varepsilon}$ denotes the error vector. In practice, the distributional features of $\boldsymbol{\varepsilon}$ are the most crucial consideration for parameter estimation, which shall be the focus of subsequent sections.

2.1. Formulation for WLS estimators

The simplest case is to consider $\bar{\varepsilon}_i$ as the annual mean of random increments from independent and identical Brownian motions with variance σ^2. Such approximations are reasonable when the assets allocated in different funds do not have substantial overlaps. In such cases, the random component of y_i becomes a Brownian motion that has evolved from time $t_{i,1}$ to $t_{i,2}$ with $y_i|x_i \sim N\left(\mathbb{E}[y_i|x_i], \sigma^2(t_{i,2} - t_{i,1})\right)$. Since $\bar{y}_i = y_i/(t_{i,2} - t_{i,1})$, $\bar{y}_i|\bar{x}_i \sim N\left(\mathbb{E}[\bar{y}_i|\bar{x}_i], \sigma^2/(t_{i,2} - t_{i,1})\right)$ and $\bar{\varepsilon}_i|\bar{x}_i \sim N\left(0, \sigma^2/(t_{i,2} - t_{i,1})\right)$. Here, errors are independent but heteroscedastic, which naturally invokes the use of WLS estimators. The obvious interpretation in such cases is that the mean performance evaluated over extended periods tend to be less volatile and thus more reliable, which in turn is allocated more weights in the WLS approach.

Derivation of WLS estimators and relevant model statistics

To calculate the WLS estimators, suppose a sample size of n, and let $\mathbf{W}$ denote the weight matrix. We then have

$$\mathrm{W}_{ii} = (t_{i,1} - t_{i,2}), \mathrm{W}_{ij} = 0, \forall i \neq j,$$

such that the WLS estimates is computed by the standard method with

$$\widehat{\boldsymbol{\beta}}_{WLS} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

and the relevant model statistics

$$\mathbb{E}\left[\widehat{\boldsymbol{\beta}}_{WLS}\right] = \boldsymbol{\beta}, \mathbb{Var}\left[\widehat{\boldsymbol{\beta}}_{WLS}\right] = \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1},$$

$$\begin{aligned} SSE_{WLS} &= \left(\mathbf{W}^{1/2}\mathbf{y} - \mathbf{W}^{1/2}\mathbf{X}\widehat{\boldsymbol{\beta}}_{WLS}\right)^T\left(\mathbf{W}^{1/2}\mathbf{y} - \mathbf{W}^{1/2}\mathbf{X}\widehat{\boldsymbol{\beta}}_{WLS}\right) \\ &= \mathbf{y}^T\mathbf{W}\mathbf{y} + \widehat{\boldsymbol{\beta}}_{WLS}^T\mathbf{X}^T\mathbf{W}\mathbf{X}\widehat{\boldsymbol{\beta}}_{WLS} - 2\widehat{\boldsymbol{\beta}}_{WLS}^T\mathbf{X}^T\mathbf{W}\mathbf{y} \\ &= \mathbf{y}^T\mathbf{W}\mathbf{y} - \mathbf{y}^T\mathbf{W}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}, \end{aligned}$$

which then permits us to estimate σ^2 with $s^2 = SSE_{WLS}/(n-d)$ with $d = 2$ and 3 for the linear and quadratic models respectively, and leads to the t-statistics for the parameter estimates as

$$t[\hat{\beta}_{i-1,WLS}] = \frac{\hat{\beta}_{i-1,WLS} - c}{s\sqrt{[(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}]_{ii}}}$$

for hypothesis tests on $\hat{\beta}_{i-1,WLS}$ with reference to c.

Computation of CI and PI for predicted annual performance

Eventually, the most important information we wish to obtain is how the fund manager will perform when we have a forecast on the overall fund market. Suppose that we set our prediction timeframe to one year, such that we want to compute the expected performance in the next year. The information above then implies that

$$\hat{y}_{new} = \mathbf{x}_{new}\widehat{\boldsymbol{\beta}}_{WLS}, \mathbb{Var}[\varepsilon_{new}] = \sigma^2,$$

where the form of $\mathbf{x}_{new}$ depends on whether the linear or quadratic model is used. We then compute the statistics for $\hat{y}_{new}$ to be

$$\widehat{\mathbb{Var}}[\hat{y}_{new}] = \mathbf{x}_{new}^T\widehat{\mathbb{Var}}[\widehat{\boldsymbol{\beta}}_{WLS}]\mathbf{x}_{new} = s^2\mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{new}$$

$$\Rightarrow \widehat{\mathbb{Var}}[y_{new}] = \widehat{\mathbb{Var}}[\hat{y}_{new}] + \widehat{\mathbb{Var}}[\varepsilon_{new}] = s^2(1 + \mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{new}).$$

The $(1-\alpha)$ CI and PI are thus constructed as

$$CI[y_{new}] = \hat{y}_{new} \pm t_{n-d,\alpha/2}s\sqrt{\mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{new}},$$

$$PI[y_{new}] = \hat{y}_{new} \pm t_{n-d,\alpha/2}s\sqrt{(1 + \mathbf{x}_{new}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_{new})},$$

where $t_{n-d,\alpha/2}$ is the upper $\alpha/2$th quantile for the t-distribution at $(n-d)$ d.f..

2.2. Derivations for GLS estimators

Nonetheless, the use of heteroscedastic errors may still be inadequate in many cases, since different funds could have substantial overlaps in assets that are correlated, such that the errors are not independent. In practice, it is impossible to quantify the extent of correlation between errors of different funds, since the exact operations of the fund are usually unknown and the fund's tactics may evolve over time. Despite such constraints, it is still possible to consider the worst case, where the errors of all funds come from the same Brownian motion. This means that whenever the operations of two funds overlap in time, over the duration of overlap their performances are perfectly correlated. In such cases, the distributional features for $\boldsymbol{\varepsilon}$ require a delicate consideration of the properties of Brownian motions, which we shall derive below.

Proposition 2.2.1: Let $B(t_1, t_2)$ and $B(s_1, s_2)$ denote two Brownian motions of variance σ^2. The covariance between $B(t_1, t_2)$ and $B(s_1, s_2)$ then becomes

$$\mathbb{Cov}[B(t_1, t_2), B(s_1, s_2)] = \sigma^2\max\big(0, \min(t_2, s_2) - \max(t_1, s_1)\big).$$

Proof: The statement holds trivially for $t_2 < s_1$, since in such cases $B(t_1, t_2)$ and $B(s_1, s_2)$ do not overlap and are thus independent. In case that overlap occurs, there are two situations to consider. Suppose that $t_1 < s_1$ and $t_2 > s_2$. We then have

$$B(t_1, t_2) - B(s_1, s_2) = B(t_1, s_1) + B(s_2, t_2),$$

where $B(t_1, s_1)$ and $B(s_2, t_2)$ are independent. We then compute variance on both sides, where

$$\mathbb{Var}[B(t_1, t_2) - B(s_1, s_2)] = \mathbb{Var}[B(t_1, s_1)] + \mathbb{Var}[B(s_2, t_2)]$$

$$\Rightarrow \sigma^2(t_2 - t_1 + s_2 - s_1) - 2\mathbb{Cov}[B(t_1, t_2), B(s_1, s_2)] = \sigma^2(s_1 - t_1 + t_2 - s_2)$$

$$\Rightarrow \mathbb{Cov}[B(t_1, t_2), B(s_1, s_2)] = \sigma^2(s_2 - s_1).$$

Suppose alternatively that $t_1 < s_1$, $t_2 < s_2$ but $t_2 > s_1$. We similarly have

$$B(t_1, t_2) - B(s_1, s_2) = B(t_1, s_1) - B(t_2, s_2)$$

where $B(t_1, s_1)$ and $B(t_2, s_2)$ are independent, such that

$$\mathbb{Var}[B(t_1, t_2) - B(s_1, s_2)] = \mathbb{Var}[B(t_1, s_1)] + \mathbb{Var}[B(t_2, s_2)]$$
$$\Rightarrow \sigma^2(t_2 - t_1 + s_2 - s_1) - 2\mathbb{Cov}[B(t_1, t_2), B(s_1, s_2)] = \sigma^2(s_1 - t_1 + s_2 - t_2)$$
$$\Rightarrow \mathbb{Cov}[B(t_1, t_2), B(s_1, s_2)] = \sigma^2(t_2 - s_1),$$

which shows that the statement in question holds for all cases.

A natural consequence of Proposition 2.2.1 is the ability to compute the covariance between ε_i and ε_j in the above-mentioned model.

Proposition 2.2.2: For $i \neq j$, we have

$$\mathbb{Cov}[\bar{\varepsilon}_i, \bar{\varepsilon}_j] = \frac{\sigma^2}{(t_{i,2} - t_{i,1})(t_{j,2} - t_{j,1})} \max\Big(0, \min(t_{i,2}, t_{j,2}) - \max(t_{i,1}, t_{j,1})\Big).$$

Proof: We note that $\mathbb{Cov}[y_i, y_j] = \sigma^2 \max\Big(0, \min(t_{i,2}, t_{j,2}) - \max(t_{i,1}, t_{j,1})\Big)$. The statement in question follows from the basic properties of covariance upon scalar multiplication, where

$$\mathbb{Cov}[\bar{\varepsilon}_i, \bar{\varepsilon}_j] = \mathbb{Cov}[\bar{y}_i, \bar{y}_j] = \mathbb{Cov}[y_i/(t_{i,2} - t_{i,1}), y_j/(t_{j,2} - t_{j,1})]$$
$$= \frac{\mathbb{Cov}[y_i, y_j]}{(t_{i,2} - t_{i,1})(t_{j,2} - t_{j,1})}$$

which then proves the statement.

With these results, we are ready to state the distributional features of $\boldsymbol{\varepsilon}$.

Proposition 2.2.3: The error vector $\boldsymbol{\varepsilon}$ follows a multivariate normal distribution denoted as $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\boldsymbol{\Omega})$, with $\Omega_{ii} = 1/(t_{i,2} - t_{i,1})$, and $\Omega_{ij} = \frac{\max(0, \min(t_{i,2}, t_{j,2}) - \max(t_{i,1}, t_{j,1}))}{(t_{i,2} - t_{i,1})(t_{j,2} - t_{j,1})}$ for $i \neq j$.

Proof: The expectations and the elements in $\boldsymbol{\Omega}$ follows directly from previous discussions on the expectations and variances of errors and from Proposition 2.3.2, so that it remains to prove the normality of $\boldsymbol{\varepsilon}$. In this case, we note that it is possible to construct a set of time durations $\{(\tau_{m,1}, \tau_{m,2})\}_{m=1}^{n}$ with no mutual overlaps, and subsequently a set of Brownian motions $\mathbf{b} = \{B(\tau_{m,1}, \tau_{m,2})\}_{m=1}^{n}$ that are mutually independent. It is always possible to discover such sets, where each error term corresponds exactly to a linear combination of Brownian motions, such that

$$\bar{\varepsilon}_i = \frac{\sum_{m=1}^{n} \chi_{(\tau_{m,1}, \tau_{m,2}) \subseteq (t_{i,1}, t_{i,2})} B(\tau_{m,1}, \tau_{m,2})}{t_{i,2} - t_{i,1}}$$

where χ denotes the indicator function. We then express the scheme above in matrix form. Construct a transformation matrix $\mathbf{A}$ of zeros and ones with $\mathrm{A}_{ij} = \chi_{(\tau_{j,1}, \tau_{j,2}) \subseteq (t_{i,1}, t_{i,2})}$, and a scale matrix $\mathbf{T}$ with $\mathrm{T}_{ii} = 1/(t_{i,2} - t_{i,1})$ and $\mathrm{T}_{ij} = 0$ for $i \neq j$. We then have $\boldsymbol{\varepsilon} = \mathbf{TAb}$. Since all elements in $\mathbf{b}$ are normally distributed and independent, $\mathbf{b}$ follows a multivariate normal distribution, such that $\boldsymbol{\varepsilon}$ must also follow a multivariate normal distribution.

With such information, we eventually arrive at the GLS estimate for the model, where

$$\widehat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Omega}^{-1}\mathbf{y}.$$

2.3. Moderations on GLS estimators

Nonetheless, the above-mentioned GLS estimators could face a number of problems when applied to actual fund performances. This is because the level of covariance assumed in the GLS estimators substantially reduces the effective amount of information available to compute the model, and in some extreme cases such computations may not even be possible. To see this, suppose that we have a collection of funds initiated and terminated at exactly the same date. Basic market intuitions teach us that the performances of such funds would at least differ to some extent, and perfect correlation is almost impossible, of which the assumptions made for the GLS estimators clearly contravene such intuitions. In addition, the $\mathbf{\Omega}$ associated with such a collection is simply a matrix of ones, which has zero determinant and is therefore not invertible, such that computation of $\mathbf{\Omega}^{-1}$ and the model cannot proceed. Even if such extreme cases are rare, there is still a possibility to have an almost noninvertible $\mathbf{\Omega}$, where parameter estimates would become unstable.

For such cases, a natural remedy is to introduce a moderation factor which uniformly shrinks the covariance between errors. Let z denote moderation factor with $z \in [0, 1]$, and let $\mathbf{\Omega}(z)$ denote the new matrix moderated by z, where we set $\Omega(z)_{ii} = \Omega_{ii}$ and $\Omega(z)_{ij} = z\Omega_{ij}, i \neq j$. Consequently, $\mathbf{\Omega}(z)$ can be viewed as a function of z, which keeps the variances but scales the covariances uniformly by the factor z. Before we proceed, it is necessary to check that $\sigma^2\mathbf{\Omega}(z)$ is indeed a covariance matrix.

Proposition 2.3.1: With $\mathbf{\Omega}(z)$ as $\Omega(z)_{ii} = \Omega_{ii}$ and $\Omega(z)_{ij} = z\Omega_{ij}, i \neq j$, $\sigma^2\mathbf{\Omega}(z)$ is a covariance matrix for all $z \in [0, 1]$.

Proof: We recall that a matrix is a covariance matrix if and only if the matrix is symmetric and positive-semidefinite. To show this, rewrite $\sigma^2\mathbf{\Omega}(z)$ as $\sigma^2\mathbf{\Omega}(z) = z\sigma^2\mathbf{\Omega} + (1-z)\sigma^2\mathrm{diag}(\mathbf{\Omega})$. Clearly, since $\mathbf{\Omega}$ is symmetric, $\sigma^2\mathbf{\Omega}(z)$ must also be symmetric, and it remains to show that $\sigma^2\mathbf{\Omega}(z)$ is positive-semidefinite. In this case, let $\mathbf{\Omega}(z)_{(K)}$, $\mathbf{\Omega}_{(K)}$ and $\mathrm{diag}(\mathbf{\Omega})_{(K)}$ denote the submatrices formed by extraction of all rows and columns indexed by $k \in K$, where we have $\sigma^2\mathbf{\Omega}(z)_{(K)} = z\sigma^2\mathbf{\Omega}_{(K)} + (1-z)\sigma^2\mathrm{diag}(\mathbf{\Omega})_{(K)}$. Obviously, $\sigma^2\mathbf{\Omega}_{(K)}$ is still a covariance matrix for all observations not indexed by K, and is therefore semidefinite with the principal minor $\left|\sigma^2\mathbf{\Omega}_{(K)}\right| \geq 0$. Similarly, $\sigma^2\mathrm{diag}(\mathbf{\Omega})_{(K)}$ is a covariance matrix for some collection of uncorrelated random variables with $\left|\sigma^2\mathrm{diag}(\mathbf{\Omega})_{(K)}\right| \geq 0$. The principal minor $\left|\sigma^2\mathbf{\Omega}(z)_{(K)}\right|$ then satisfies the inequality

$$\left|\sigma^2\mathbf{\Omega}(z)_{(K)}\right| \geq \left|z\sigma^2\mathbf{\Omega}_{(K)}\right| + \left|(1-z)\sigma^2\mathrm{diag}(\mathbf{\Omega})_{(K)}\right|$$
$$\Rightarrow \left|\sigma^2\mathbf{\Omega}(z)_{(K)}\right| \geq z\left|\sigma^2\mathbf{\Omega}_{(K)}\right| + (1-z)\left|\sigma^2\mathrm{diag}(\mathbf{\Omega})_{(K)}\right|.$$

Since $z \in [0, 1]$, we have $z \geq 0$ as well as $1 - z \geq 0$, which implies that $\left|\sigma^2\mathbf{\Omega}(z)_{(K)}\right| \geq 0$. Since the index set K is arbitrary, the result applies to all principal minors, such that $\sigma^2\mathbf{\Omega}(z)$ must be positive-semidefinite. Combined with the symmetric property, $\sigma^2\mathbf{\Omega}(z)$ is a covariance matrix.

This result ensures that whenever we moderate $\mathbf{\Omega}$, the moderation procedure always creates $\mathbf{\Omega}(z)$ as a valid covariance matrix amenable to computations for the GLS estimators. The moderated estimators are thus a function of the moderation factor z, where we have

$$\widehat{\boldsymbol{\beta}}_{GLS}(z) = (\mathbf{X}^T\mathbf{\Omega}(z)^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{\Omega}(z)^{-1}\mathbf{y}$$

with a set of possible estimates over the continuum $z \in [0, 1]$. It is not hard to see that $z = 0$ corresponds to the WLS estimators, and $z = 1$ corresponds to the unmoderated GLS estimators, such that $\widehat{\boldsymbol{\beta}}_{GLS}(z)$ covers all cases intermediate between WLS and GLS estimators.

2.4. General procedures for model selection and sensitivity analysis

The above-mentioned results mean that we can choose different values of z to compute a sequence of $\widehat{\boldsymbol{\beta}}_{GLS}(z)$, in order to evaluate the sensitivity of parameter estimates to correlations between errors. With these, we are then ready to state a complete procedure for the analysis of fund performance as below.

1. Transform fund performances to logarithmic scale and compute their arithmetic means.
2. Fit linear and quadratic models for the transformed data with WLS estimators. Fit also a null model for reference.
3. Select the model with the smallest BIC.
4. With the selected model, choose a set of moderation factors to compute a set of GLS estimators.
5. Compare the results from GLS estimators with that of the WLS estimators to assess model sensitivity. If sensitivity is detected, state the critical level of correlation beyond which the model is not reliable.
6. Obtain bands for CI and PI to assess fund performance in different market situations.

With all the mathematical details discussed, the sections below will perform a case study for some top fund managers in the Chinese fund market.

Section 3. A preliminary case study on top performers in the Chinese fund market

To commence, we browse for publicly accessible information on the performances for all fund managers in the Chinese fund market. A complete list could be obtained on Sina Finance, where overall performance is measured on the cumulative performance of a fund manager relative the mean performance of the overall fund market. For the purpose of this write-up, we extract relevant data delineated at the start of Section 2 from the list of the top 40 performers, with the additional restriction that each selected person must have experience with at least 10 funds. This is to ensure an adequate amount of historical information to estimate both the linear and the quadratic models, where we follow the rule of at least 5 data points per predictor term. Model selection and analysis is then performed by the procedure delineated in Section 2.4. The sections below will tabulate the results for all included data, followed by a detailed analysis of two selected fund managers to demonstrate the principles.

3.1. Overall results for model selection and sensitivity analysis

Within the 40 top performers, 8 has met the requirement for experience in at least 10 funds. The procedures for model selection, model preference, as well as the results for sensitivity analysis are summarized in Table 3.1.1, where BIC_0, BIC_1 and BIC_2 correspond to the null, linear and quadratic models respectively. The set of z values includes $z = 0$, 0.2, 0.4, 0.6, 0.8 and 1, of which a threshold value is provided to denote the maximum z below which the GLS estimators remain stable about the WLS estimators.

Table 3.1.1: Overall model selection and sensitivity analysis

Name	BIC_0	BIC_1	BIC_2	Preference	Threshold z
Li Wei	-10.83	-16.21	-13.93	Linear	0.8
Sun Wei	-8.38	-19.52	-17.56	Linear	0.8
Tao Can	-1.87	-10.02	-11.88	Quadratic	1.0
Chen Hao	-6.00	-20.67	-25.17	Quadratic	0.8*
Xiao Nan	-18.26	-15.97	-13.83	Null	1.0
Chen Zhou	-15.65	-28.79	-26.70	Linear	0.8
You Linfeng	-29.10	-26.98	-24.92	Null	1.0
Wang Zonghe	-22.21	-24.17	-22.30	Linear	0.2*
*: Noninvertible $\mathbf{\Omega}(z)$ for $z = 1$.					

Generally, we can see that the WLS estimators do indicate a diversity of performance trends for different fund managers. In terms of robustness of estimates, the WLS estimators usually remain stable even in the presence of substantial correlation between errors, which holds for most cases for $z \leq 0.8$. Nonetheless, it is still possible for some WLS estimates to be unstable even at relatively low value of z, such that interpretation of the model should proceed cautiously. For $z > 0.8$, most GLS estimators start to deviate substantially from the WLS estimators, and in some cases the error covariance matrix becomes almost noninvertible. In actual practice, the null, linear and quadratic models naturally invoke different interpretations upon different market situations.

The null model indicates a relatively constant performance over different market situations. To determine whether such a performance deviates from zero return, we could do t-tests on the intercept term with reference to $\bar{y} = 0$, which is the point of zero return when transformed back to the linear scale. Nonetheless, it is important not to extrapolate this constant performance over market extremes not covered in the current data, since it is still possible for such funds to incur unforeseeable losses amid an economic downturn.

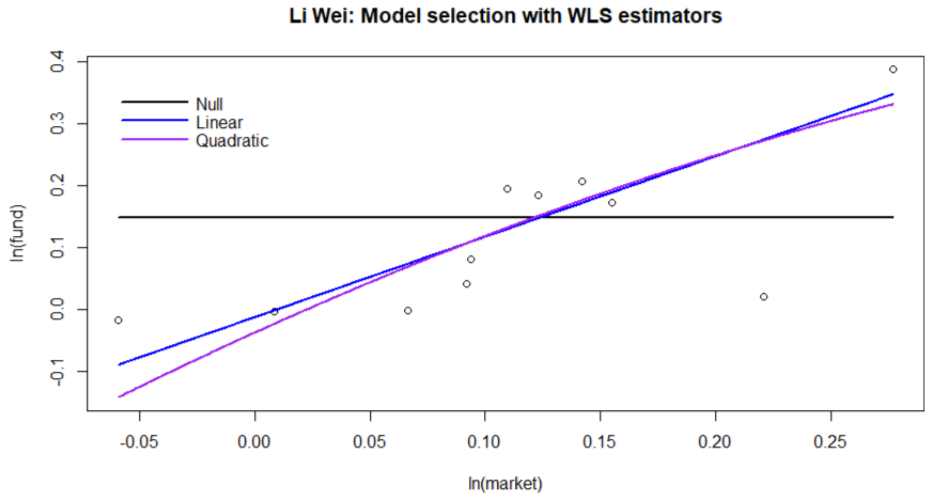
Compared to the null model, the linear model indicates a performance trend that correlates with the market situation. To determine whether performance is more or less volatile than the overall fund market, we could do t-test on the slope term with reference to $\beta_1 = 1$. To determine whether there is a systematically superior performance over the market, we could do t-test on the intercept term with reference to $\beta_0 = 0$. Prediction of future performance should only be done when we already have a reasonable forecast on the overall market.

Compared to the linear model, the quadratic model indicates a nonlinear relationship between the fund manager's performance and the overall market, which cannot be described by simple rules, but must be examined on a case-to-case basis. We proceed to discuss two of the cases collected to demonstrate the basic principles for inference and prediction.

3.2. Case 1: Li Wei

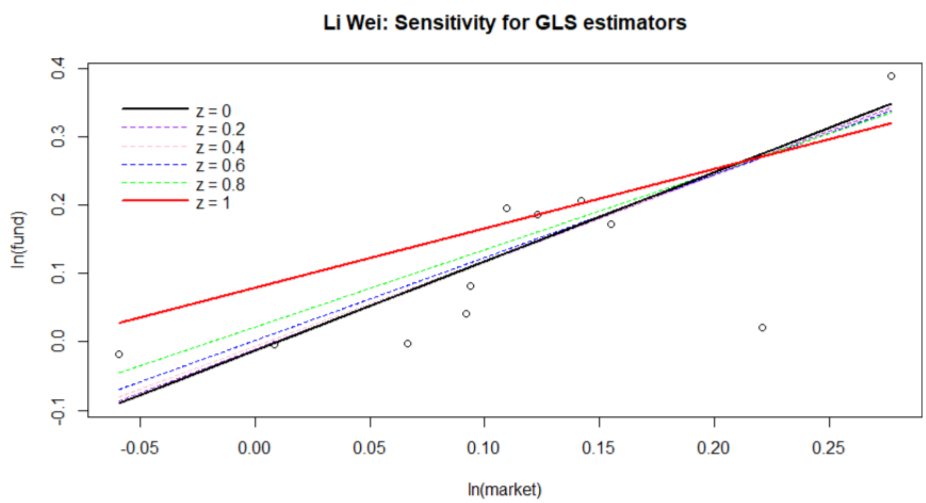
Li Wei is the 7th top performer with 571.20% return over the course of his career relative to 250.69% return for the benchmark. He currently has experience in 11 funds. To reiterate the results, the null, linear and quadratic models based on WLS estimators are displayed in Plot 3.2.1.

Plot 3.2.1: Models for historical annual performance of Li Wei's funds



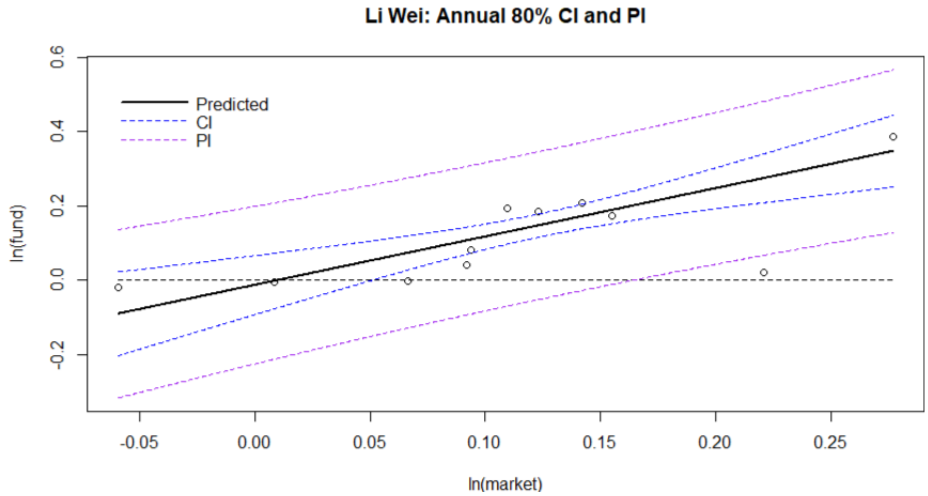
We recall that in the tabulated results, the linear model is selected for this case, so that we proceed with relevant hypothesis tests for the parameter estimates. For the intercept term, the p-value is 0.824 for $H_0: \beta_0 = 0$ over $H_A: \beta_0 \neq 0$, which indicates that Li Wei's funds do not appear to systematically outperform the market. Hypothesis test for $H_0: \beta_1 = 1$ over $H_A: \beta_1 \neq 1$ yields a t-statistic of 0.7049 on 9 d.f., which is rather mediocre. The combined results indicate that Li Wei's funds have comparable mean annual performances with the market. Subsequent sensitivity analysis shows that the GLS estimators remain stable about the WLS estimators for $z = 0$, 0.2, 0.4, 0.6, 0.8, but not for $z = 1$, as can be seen from Plot 3.2.2.

Plot 3.2.2: Sensitivity of GLS estimators for Li Wei's funds



Eventually, the bands for 80% CI and PI are displayed in Plot 3.2.3. The result shows that when market return is above 17% $\approx e^{0.16} - 1$, the PI does not include the line of zero return, such that the positive return for Li Wei's funds is significant at 80% level. Of course, if we are more risk-averse, we could construct wider and thus more conservative CI and PI.

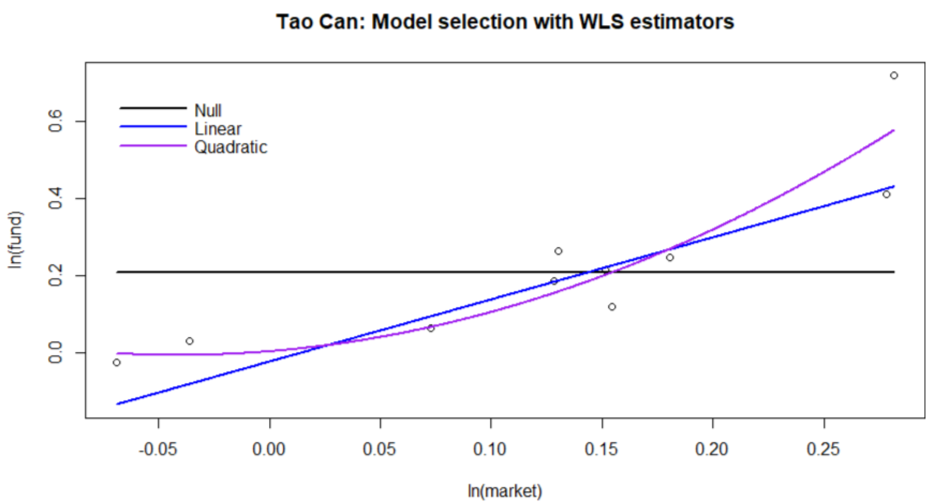
Plot 3.2.3: 80% CI and PI on predicted returns for Li Wei's funds



3.3. Case 2: Tao Can

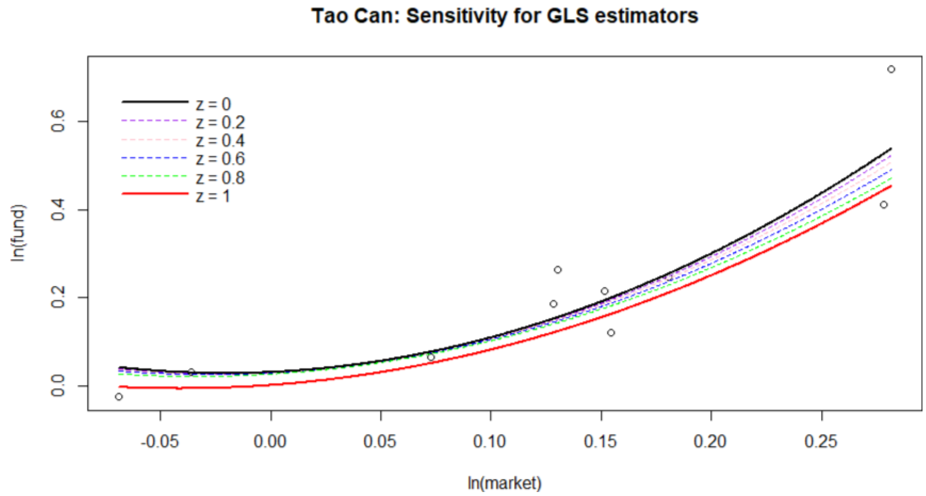
Tao Can is the 9th top performer with 548.10% return over the course of his career relative to 288.37% return for the benchmark. He currently has experience in 10 funds. The null, linear and quadratic models based on WLS estimators are similarly displayed in Plot 3.3.1.

Plot 3.3.1: Models for historical annual performance of Tao Can's funds



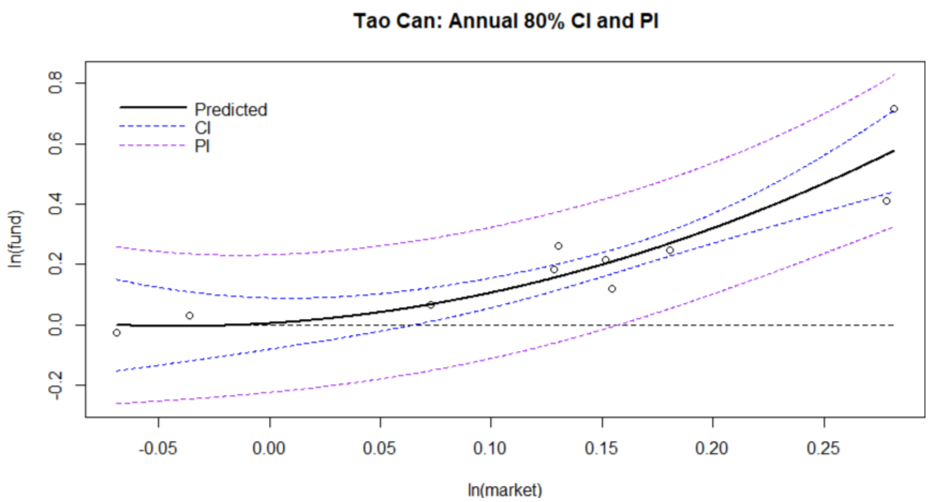
In Tao Can's case, we recall that the quadratic model is preferred. The results for sensitivity analysis are then displayed in Plot 3.3.2. Generally, increased error correlations tend to yield slightly more conservative predictions from the GLS estimators, while the functional form of the performance trend remains stable.

Plot 3.3.2: Sensitivity of GLS estimators for Tao Can's funds



Eventually, the 80% CI and PI are displayed in Plot 3.3.3, where significantly positive fund return occurs for market returns above 16% $\approx e^{0.15} - 1$.

Plot 3.3.3: 80% CI and PI on predicted returns for Tao Can's funds



Conclusion

This small write-up has derived some useful theoretical formulations for WLS and GLS estimators, and proposed suitable protocols to select appropriate models and evaluate their robustness at different levels of error correlations. A preliminary study performed on the Chinese fund market has delineated the basic principles on which predictions on future fund performances could be built to aid relevant investment decisions. Nonetheless, due to the sparsity of available information, we must also concede that the small sample size necessarily yields conservative recommendations, and model assumptions on the proposed normality and covariance structure between errors cannot be tested conclusively. Future directions include more comprehensive analysis of relevant fund markets, which would then provide more data to evaluate the validity and generalizability of the proposed models.