PINNACLE INVESTMENT RESEARCH
Summer Internship

Breaking Down Bias Statistic to Systematic and Unsystematic Risk

Pinnacle Investment
Jakarta, DKI Jakarta, Indonesia

Indrik Wijaya
NUS Applied Mathematics and Economics
A0107380M
July 1, 2016

TABLE OF CONTENTS

EXECUTIVE SUMMARY

This report studies whether it is possible to break down bias statistic into the common risk components: Systematic Risk and Unsystematic Risk. Bias statistic helps to determine the accuracy of risk forecast of a certain portfolio model currently pursued and developed by the Research and Investment group at Pinnacle Investment. If we can break it down, we can determine whether Systematic Risk and/or Unsystematic Risk has a higher contribution in causing a model to underpredict or overpredict risk.

To determine whether this is possible, a simulation over wide range of values is done on the equation that relates Bias Statistic with a measure of Systematic Risk and Unsystematic Risk. The report argues that we can indeed break down the Bias Statistic into the two distinct risk components. Knowing this fact allows us to improve the accuracy of our risk forecast, and hence, improving our investment competitive strategy.

1.0 INTRODUCTION

Based on Factor-based pricing models, we can write our portfolio return as a summation of its two distinct sources: Systematic and Unsystematic (or specific).

$$R = \underbrace{\sum_k \beta_k f_k}_{\text{systematic}} + \underbrace{\epsilon}_{\text{specific}},$$

where β_k is the exposure of the stock to factor k, f_k is the factor return, and ϵ is the specific return. This is also known as the *Arbritrage Pricing Theory* (APT). For simplicity, we will write our return at any time, t as

$$R = \underbrace{S_t}_{\text{systematic}} + \underbrace{U_t}_{\text{unsystematic}} \tag{1.0.1}$$

Similarly, we can write the portfolio variance, assuming that factor returns are uncorrelated with the systematic returns and specific returns are uncorrelated among themselves as

$$\mathrm{Var}(R_P) = \sum_k \sum_l \beta_k^P F_{kl} \beta_l^P + \sum_n w^2 \mathrm{Var}(\epsilon)$$

where F_{kl} is the predicted covariance between factors k and l and w is the weight of portfolio. Again, for simplicity we will write our variance as

$$\mathrm{Var}(R_P) = \underbrace{\sigma_{st}^2}_{\text{systematic}} + \underbrace{\sigma_{ut}^2}_{\text{unsystematic}} \tag{1.0.2}$$

1.1 Single-Window Bias Statistic

Conceptually, **bias statistic** represents the ratio of realized risk to forecast risk. Let R_t be the return to our portfolio over period t, and let σ_t be the beginning-of-period volatility forecast. In order to evaluate the **Bias Statistic**, we need to first determine *standardized return*. Assuming perfect forecast, it can be written as

$$b_t = \frac{R_t}{\sigma_t},$$

which has an expected standard deviation of 1. The **bias statistic** for our portfolio is then the *realized standard* deviation of standardized returns,

$$B = \sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(b_t - \bar{b})^2} \tag{1.1.1}$$

where T is the number of periods in the observation window.

Assuming normally distributed returns and perfect risk forecasts, for sufficiently large T, B is approximately normally distributed about 1, and roughly 95% of the observations fall within the confidence interval,

$$B \in [1 - \sqrt{2/T}, 1 + \sqrt{2/T}]$$

If B_n falls outside this interval, we reject the null hypothesis that the risk forecast was accurate. If returns are not normally distributed, however, fewer than 95% of the observations will fall within the confidence interval, even for perfect risk forecasts.

1.2 Rolling-Window Bias Statistics

Let T be the length of the observation window, which corresponds to the number of months in the sample period. One possibility is to select the entire sample period as a single window, and to compute the bias statistic as in Equation (1.1.1). This would be a good approach if financial data were stationary, as sampling error is reduced by increasing the length of the window. In reality, however, financial data are not stationary. It is possible to significantly overpredict risk for some years, and underpredict it for others, while ending up a bias statistic close to 1.

Often, a more relevant question is to study the accuracy of risk forecasts over 12−month periods. For this purpose, we define the rolling 12−month bias statistic for portfolio n,

$$B^\tau = \sqrt{\frac{1}{11}\sum_{t=\tau}^{\tau+11}(b_t - \bar{b})^2},$$

where τ denotes the first month of the 12−month window. The 12−month windows are rolled forward one month at a time until reaching the end of the observation window. If T is the number of periods in the observation window, then each portfolio will have $T - 11$ (overlapping) 12−month windows.

It is useful to consider, for a collection of N portfolios, the mean of the rolling 12−month bias statistics,

$$\overline{B^\tau} = \frac{1}{N}\sum_n B_n^t.$$

Note that the observations in a rolling scenario are no longer independent (and are, in fact, autocorrelated). Assuming normal distributions and perfect risk forecasts, the autocorrelation-adjusted 95% confidence interval is $\left[\mathbf{0.66}, \mathbf{1.34}\right]$. $B^\tau < 0.66$ (resp. > 1.34) means that the particular risk forecast is underpredicting (resp. overpredicting) the risk.

When we carry out the bias statistic test, we will get the percentage of our forecasts that underpredict and overpredict the risk. This report is useful when the percentage of under-prediction and/or overprediction is high. Even though a large majority of our forecasts may fall in the confidence interval, the percentage of those lying outside the confidence interval may also be high. Thus, this report aims to break down the bias statistic equation into the two distinct risk components, which will allow us to determine which component causes the problem.

2.0 ANALYSIS

From the definition of bias statistic, we know that we are basically evaluating the realized standard deviation of standardized returns, ie $SD(b_t)$. This is equivalent to evaluating $\mathrm{Var}(b_t)$. Our aim is then to break this down into systematic risk ($\mathrm{Var}(\mathrm{systematic\,return})$) and unsystematic risk ($\mathrm{Var}(\mathrm{unsystematic\,return})$). If we can successfully do this, we will be able to determine the contribution of each risk component.

2.1 Mathematical Outline

We want to derive an equation that relates $\mathrm{Var}(b_{nt})$ with ($\mathrm{Var}(\mathrm{systematic\,return})$) and ($\mathrm{Var}(\mathrm{unsystematic\,return})$). We begin with writing our portfolio return and variance using Equation (1.0.1) and Equation (1.0.2). Then, we can write our standardized return, b_{nt} as $\frac{S_t+U_t}{\sqrt{\sigma_{st}^2+\sigma_{ut}^2}}$. Assuming

$$\frac{S_t + U_t}{\sqrt{\sigma_{st}^2 + \sigma_{ut}^2}} = x\frac{S_t}{\sigma_{st}} + y\frac{U_t}{\sigma_{ut}} \tag{2.1.1}$$

where x and y is a constant to be determined, we can then write $\mathrm{Var}(b_t)$ as

$$\begin{aligned}\mathrm{Var}\left(\frac{S_t + U_t}{\sqrt{\sigma_{st}^2 + \sigma_{ut}^2}}\right) &= \mathrm{Var}\left(x\frac{S_t}{\sigma_{st}} + y\frac{U_t}{\sigma_{ut}}\right)\\ &= x^2\mathrm{Var}\left(\frac{S_t}{\sigma_{st}}\right) + y^2\mathrm{Var}\left(\frac{U_t}{\sigma_{ut}}\right)\end{aligned}$$

By determining the value of x^2 and y^2, we will be able to determine the contribution of each risk component in our risk model. Before determining the value of x and y, we need to carry out a simulation over wide range of values of the dependent variables: S_t, σ_{st}, U_t and σ_{ut}, in order to determine whether 2.1.1 holds. To do this, we want to minimize $f(x)$ where

$$f(x) = \left(\frac{S_t + U_t}{\sqrt{\sigma_{st}^2 + \sigma_{ut}^2}} - x\frac{S_t}{\sigma_t} - y\frac{U_t}{\sigma_{ut}} \right)^2$$

To ascertain that 2.1.1 holds, we want $f_{\min}$ to be 0 or as close as possible to 0. If f_{min} is far greater than 0, it means that the equation does not hold and hence, we cannot break down the **bias statistic** into the two risk components. This report will not explain in detail of this step as it is not the main focus of our objective. In fact, this result holds for most values ($> 99\%$) that returns and standard deviation of a portfolio usually take in the real world. Our main objective is to determine whether this relation holds when we take the variance on both sides.

2.2 Procedure

This section outlines the simulation procedure, which uses **multiple linear regression**. Our aim is to determine the absolute difference between $\mathrm{Var}(b_t)$ and Var(predicted bias), where predicted bias is obtained from multiple linear regression of $\frac{S_t+U_t}{\sqrt{\sigma_{st}^2+\sigma_{ut}^2}}$ with $\frac{S_t}{\sigma_{st}}$ and $\frac{U_t}{\sigma_{ut}}$. The parameters used to generate the normal distribution for σ_{st} and σ_{ut} are: $\mu_{\text{sys}} = 0.2$; $\sigma_{\text{sys}} = 0.1$; $\mu_{\text{unsys}} = 0.1$; $\sigma_{\text{unsys}} = 0.02$. These values are kept constant throughout the simulation.

Furthermore, the parameters used to generate the normal distribution for predicted return S_t and U_t are: μ(bias sys), $\sigma_{st} * \sqrt{1 + \mathrm{E}[\text{bias systematic}]}$, μ(bias unsys), $\sigma_{ut} * \sqrt{1 + \mathrm{E}[\text{bias unsystematic}]}$. These parameters will be our dependent variables. We carry out 50 different iterations for each set of dependent variables, in which each simulation uses 4000 randomly generated data that are representative of real world's values for return and standard deviation. 50 was picked as higher values do not improve our results significantly. The summary of the simulation can be found in the next page.

Simulation Procedure

Step 1. Establishing Equation 2.1.1

1. Vary one or more dependent variables and keep the rest constant (the constant value is kept at 0.15 and 0.20 for return and standard deviation, respectively).
2. Find f_{min} and record any set of values that violate the equation. We use 0.1% tolerance level to accept whether our result is accurate.
3. Repeat Step $1-4$ for different variations of S_t, U_t, σ_{st} and σ_{ut}.

Step 2. Find Predicted Bias

1. Generate randomly and normally distributed 4000 values for each of $S_t, U_t, \sigma_{st}, \sigma_{ut}$.
2. Carry out multiple linear regression for $\frac{S_t+U_t}{\sqrt{\sigma_{st}^2+\sigma_{ut}^2}}$ with $\frac{S_t}{\sigma_{st}}$ and $\frac{U_t}{\sigma_{ut}}$. Take note of the regression coefficients as x, y and z, which corresponds to $\frac{S_t}{\sigma_{st}}$, $\frac{U_t}{\sigma_{ut}}$ and the constant respectively.
3. Evaluate the predicted bias for each set of values, which is equal to $x\frac{S_t}{\sigma_{st}} + y\frac{U_t}{\sigma_{ut}} + z$. We can observe that each value of predicted bias is approximately equal to the true value.

Step 3. Find the difference between actual bias and predicted bias

1. Evaluate the variance of $\frac{S_t+U_t}{\sqrt{\sigma_{st}^2+\sigma_{ut}^2}}$, of $\frac{S_t}{\sigma_{st}}$, of $\frac{U_t}{\sigma_{ut}}$ and of predicted bias for all the randomly and normally distributed 4000 values.
2. Evalute the difference between $\mathrm{Var}(b_t)$ and Var(predicted bias). We use 7.5% tolerance level to accept whether our result is accurate. The amount of accepted value will be recorded as **Acceptance Rate**.

Step 4. Break down the bias statistic

1. Perform multiple linear regression for $\text{Var}\left(\frac{S_t+U_t}{\sqrt{\sigma_{st}^2+\sigma_{ut}^2}}\right)$ with $\text{Var}\left(\frac{S_t}{\sigma_{st}}\right)$ and $\text{Var}\left(\frac{U_t}{\sigma_{ut}}\right)$.
2. The regression coeffiecients will be the weight for systematic component (α) and unsystematic component (β).
3. $\alpha\text{Var}\left(\frac{S_t}{\sigma_{st}}\right)$ and $\beta\text{Var}\left(\frac{U_t}{\sigma_{ut}}\right)$ will be the contribution of systematic risk and unsystematic risk respectively.

2.3 Empirical Results

In this report, we will not include results for the simulation to establish Equation 2.1.1. We only note the results of our main focus which are **acceptance rate**, weight for systematic risk (α) and weight for unsystematic risk (β) The regression is done repeatedly in MATLAB. The important results are summarized below:

Table 1: Acceptance Rate, α and β for different sets of dependent variables

E[Sys]	E[Unsys]	Mean(Sys)	Mean(Unsys)	Accept(5%)	Accept(10%)	α	β
0.2	-0.1	0	0	73%	100%	0.7624	0.3971
0.1	-0.1	0	0	74%	100%	0.5958	0.1465
0.1	-0.1	0.1	0	90%	100%	0.5691	0.3582
0.2	-0.2	0	0	100%	100%	0.6743	0.3651
0.2	-0.2	0	0.05	100%	100%	0.7632	0.3485
0.2	-0.2	0	0.1	66%	100%	0.7957	0.2865
0,1	-0.1	0.1	0.05	48%	98%	0.5002	0.3768
0.2	0.1	0	0	16%	80%	0.6237	0.3223
0.1	-0.2	0	0	18%	86%	0.7806	0.3628

*other results are still pending

From **Table 1**, we can clearly see that the **acceptance rate** is high for tolerance value of 10% for all simulations. The **acceptance rate** is lower for tolerance value of 5%, yet is still high generally. Further simulation proves that this break down holds for E[bias systematic], μ(bias systematic) $\geq$

0 and E[bias unsystematic], μ(bias unsystematic) ≤ 0 for even tolerance value of 5%. If either inequality in the relations is reversed, the difference between Var(b_t) and Var($predictedbias$) will be around $10\% - 25\%$, which is still relatively small. Hence, we can still use this break down but we need to consider other models since the break down is not as accurate.

Here is a sample of one simulation which is done over 4000 data points over 50 iterations. We can clearly see that the value of Var(b_{nt}) is close to Var()predicted bias). **TABLE IS STILL PENDING**

3.0 CONCLUSION

The simulation appears to be promising as it produced the desired result, which is that we can break down the bias statistic equation into the two risk components: systematic risk and usnsytematic risk. Since the simulation spans through a wide range of possible values for return and standard deviation that are representative of the actual values, we can use the results for further analysis on the contribution of each risk component. Nonetheless, since the method used is multiple linear regression, there are certain limitations to this model. We might need to use other methods (if there is any) to further complement our findings. Yet, we can still utilize this model as long as our data do not violate the assumptions needed for multiple linear regression to be accurate.