Monthly Measurement of Daily Timers

William N. Goetzmann  
*Yale School of Management*

Jonathan Ingersoll Jr.  
*Yale School of Management*

March 11, 1998

Abstract: The Henriksson-Merton (1981) parametric measure of timing skill depends crucially on the assumption that timers make irrevocable decisions over the course of the period of return measurement. We examine the size and power of the The Henriksson-Merton test through simulations of a market and a timer who can move in and out of the market on a daily basis. Our simulations show that when the Henriksson-Merton measure is applied to the monthly returns of a timer who makes choices daily, the measure is very weak. The probability of rejecting the null when the timer has perfect foresight is less than 50%. Henriksson-Merton (1981) also show that a timer implicitly provides a put on the market to the client. The reason for the weakness of the test using monthly data is that the *ex post* realized value of the monthly put is a poor approximation to the cumulated value of a sequence of daily puts over the same month. This mis specification is different than the limitations described by Glosten & Jagannathan (1994) but similar in effect -- a tradeoff between timing vs. selection values are identified. We estimate magnitude of the implicit put provided by the timer and find that the use of monthly data biases it downward. We propose a simple solution that alleviates the problem without collecting daily timer returns. Our solution uses daily returns to an index correlated to the timer’s risky asset. We cumulate the value of a daily put over each month to form a regressor to capture timing skill. Our corrected HM measure applied to monthly returns is much more powerful and reduces the bias in the estimated put value.

For a current copy of the paper, please contact:

William N. Goetzmann  
*Yale School of Management*  
Box 208200  
New Haven, CT 06520-8200  
william.goetzmann@yale.edu  
http://viking.som.yale.edu

Acknowledgments:  
We thank Don Chance, Steve Shellans and Ed Owens for useful discussions on this topic. All errors are the sole responsibility of the authors.
**Introduction**

Henriksson-Merton (1981) [HM] develop a logically appealing measure of market-timing skill. Their analysis is based upon the simple intuition that a market-timer effectively provides a put to the client. When the market is up, the perfect timer is fully invested in the risky asset. When the market is down, the perfect timer will be holding the riskless asset. HM show how a simple parametric test -- a regression of portfolio returns on two variables -- can be used to estimate a manager’s timing skill. In this paper, we examine the size and power of the HM parametric test. We focus on the problem of using the test on monthly data when managers make daily timing decisions. Our simulations show that when the Henriksson-Merton measure is applied to the monthly returns of a timer who makes choices daily, the measure is strongly biased against detecting timing skill. In particular, the measure under-estimates the magnitude of the implicit put provided by the manager. We propose a simple solution that alleviates the problem without collecting daily manager returns. Using an instrument for the daily risky asset returns, we show that a more accurate estimate of the timer’s put can be calculated. This variable dramatically reduces the bias in the HM measure and increases the power of the test.

Several researchers have studied the Henriksson-Merton timing measure.\(^1\) Glosten and Jagannathan [GJ] show it to be a special case of a more general contingent claims approach to performance evaluation. They propose an improved version of the HM test that allows for managed portfolios to represent bundles of multiple options with different strikes. Implicit in the empirical

\(^1\)Grinblatt and Titman (1989) and Jagannathan and Korajczyk (1986) show HM can confuse timing skill with a option strategies. Admati, Battacharya, Pfleiderer and Ross (1986) point out that the functional form of the HM measure is limiting, since variation in the intensity of the timing signal will make the timer’s payoff non-linear in market returns. Other studies, for example, Lehman and Modest (1987) use a non-linear form for estimation for this reason.
application of the GJ framework, however is the presumption that the options share a common maturity. This is the heart of the problem we address. Not only is the timer effectively holding or mimicking a bundle of options with varying strikes, these options are effectively being rolled over at a frequency potentially unequal to the interval of return estimation. Only one paper to our knowledge points out the magnitude of this monthly data problem. Chance and Hemler (1998) strongly reject the null of no timing ability for a manager using daily data but find that all evidence of timing ability disappears when monthly data are used. They argue that using monthly data essentially implies that most standard timing tests are mis-specified and thus it should come as little surprise that few researchers to date have found evidence of timing ability by professional managers.

In general, evidence on the ability of investment managers to time the market is mixed. Studies of mutual fund timing skill such as Merton (1981), Merton and Henriksson (1981), Chang and Lewellan (1984), Treynor and Mazuy (1966), Henriksson (1984) Grinblatt and Titman (1989) generally find little evidence of timing skill. On the other hand, Ferson and Schadt (1996) find some evidence of manager timing skill when macroeconomic conditions are accounted for, and Graham and Harvey (1997) find evidence of timing skill using certain benchmarks. Wagner, Shellans and Paul (1992), Brocanto and Chandy (1994) and Chance and Hemler (1998) all find some positive timing evidence as well. Brown, Goetzmann and Kumar (1997) find evidence that the Dow Theory worked as a timing strategy. While our study focuses specifically on a correction for the Henriksson-Merton parametric test of timing skill, it may be that at least some of the ambiguity in results is due to the fact that most studies rely upon monthly returns. It also has direct implications for the GJ test. Allowing for more frequent timing activity may improve on the power of the tests they propose for identifying manager skill in general -- not simply improving timing tests.
This paper is organized as follows. The next section describes the Henriksson-Merton parametric test of timing skill. Section 3 describes our simulations and section 4 presents the results. Section 5 concludes.

I. The Henriksson-Merton Parametric Test of Timing Skill

In their 1981 paper, Henriksson and Merton develop two tests of timing skill. One is a non-parametric test which relies upon knowing the timer’s forecast of the market. The other is a test that relies solely on the returns generated by the timer. For cases when the timer’s forecast is known, the non-parametric test is a direct test of the timer’s forecasting skill. In many circumstances, the timer’s forecast is unknown. Few investment managers report their forecasts as well as their performance. Tests of the timing ability of mutual fund managers, for example, typically rely upon monthly fund returns.

The Henriksson-Merton parametric test is a linear regression of the timer’s excess returns on

\[ R_{t} = \alpha + \beta R_{m,t} + \gamma \max \{-R_{m,t}, 0\} + \epsilon_{t} \]

two variables.

The first variable is excess return of the risky asset (i.e. the market) and the second variable capturing the value of the implicit put. It takes on the value 0 when the excess return of the market is positive and it exactly offsets losses when the market drops. A perfect timer should have a market

\[ \text{Pesaran and Timmerman (1994) show that the non-parametric test is equivalent to a Fisher’s exact test about a 2X2 matrix.} \]
\( \beta \) of 1 and a timing \( \gamma \) of one. This corresponds to a long position in the equity asset and a long position in a put with a maturity of one period, which is struck at-the-money at the beginning period price. Notice that this formulation implicitly assumes that the timer with perfect foresight about next period’s risky asset premium will either be in the market over the entire period, or out of the market over the entire period. In fact, if a perfect timer could trade within the period, then equation 1 is mis-specified. In particular, the variable capturing the value of the implicit put does not account properly for the value of intermediate investment decisions. For example, suppose a money manager could trade daily. Even in an up month for the market, the daily timer should have a positive return. This would not affect the estimated timing coefficient \( \gamma \). Instead it would affect the market coefficient \( \beta \) estimate or the intercept term. In fact, this example demonstrates that any positive excess return achieved by the timer in an up month for the market will not be credited. This problem is exacerbated as the difference between the decision horizon and the evaluation horizon grows. Many investment managers report only quarterly performance. As the horizon grows, the frequency of negative period returns for the risky asset decreases, and so does the power of the HM parametric test, which relies upon the covariance of the manager returns with the put value conditional upon the risky asset underperforming. Consequently, as the below simulations will show, the HM test for daily timers using monthly data is extraordinarily weak.

The best solution to the problem is to collect data that corresponds to the frequency with which the timer makes decisions. This is typically not possible. While Busse (1997) collects daily mutual fund data for investigating whether mutual fund managers in general time the variance of the market and Chance and Hemler (1998) have daily data for a limited number of market-timers, almost no money manager reports daily results in a form generally accessible to researchers and analysts.
An alternative to collecting daily data is to collect daily data on the risky asset alone. Daily S&P 500 returns, for example, can be used to construct an instrument that is correlated to the daily put values. More precisely, we cumulate the value of the daily puts over the month to estimate the monthly value of a daily timer’s skill.

\[ R_t = \alpha + \beta R_{mt} + \gamma \prod_{\tau=t-1}^{t} \max[1-R_{m\tau}, 1] + \epsilon_t \]

Even when the risky asset used by the timer is not available, as long as it is correlated, this specification provides an improvement. Thus, we substitute the value of a monthly put on the market with a rolling account through the year of the gains to having a sequence of daily market puts.

II. Simulations

We conduct simulations to examine the performance of the HM parametric test using daily returns, monthly returns with the monthly put values and monthly returns with the cumulated daily put values. We then report the median values for HM coefficients as well as the frequency with which the null is rejected for differing levels of timing skill.

II.1 Simulating the market

For each simulation, we generate ten years (2,520 days) of daily excess returns to the risky asset as i.i.d random variables with an annualized mean of 10% and an annualized standard deviation of 16%. These are exponentiated (after appropriate correction of the mean) to create log-normal excess returns.

II.2 Simulating the timer
We define perfect timing skill as the ability to forecast the excess return on the next day with no error. The simulated timer takes a 100% position in the risky asset on days for which the forecast is positive and is in the riskless asset, yielding zero excess return, on days when the forecast is negative. It is also useful to define imperfect timing skill in this context. Intermediate skill levels are generated by imperfect forecasts as mixtures of the true daily value with a random daily excess return from the ten years sample. For example, a skill level of .5 indicates that the timing forecast excess return for day $t$ is $\frac{1}{2}R_{m,t} + \frac{1}{2}R_{m,k-t}$. Skill levels range from no foresight to perfect foresight. The no-foresight skill level is equivalent to the timer completely knowing the 10-year sample distribution of returns, but not being able to distinguish one day from another. One potential concern with this approach is that the variance of the imperfect forecast of the excess return is lower than the variance of either the perfect timer or the no foresight timer. This is because the mixing procedure effectively creates a portfolio. There are alternative approaches to parameterizing timing skill based upon random draws, and we are currently evaluating their relative advantages. It is not obvious how the reduction in forecast variance would affect our results, however.

III. Results

III.1 Median Values

Table I reports the median estimated values for the coefficients $\alpha$, $\beta$ and $\gamma$ over 500 simulations of the market for each skill level and for each of the three HM test specifications. The bottom panel reports the median coefficients for the timing variable. Notice that the monthly
sampled data has a strong downward bias in the timing coefficient. A perfect timer should have a coefficient of 1, but the median value in the table is .17. This downward bias is consistent with the fact that the effective put is measured with error. The monthly test using the cumulated timing variable is unbiased for the perfect timer. It is interesting to note that even using daily data, the HM test yields downward-biased timing coefficients for modest levels of timing skill. Since we would expect to find timers with anything other than modest ability (due to the efficient market theory) this is troublesome evidence. Even with perfect data the estimated put is below the true value.

The magnitude of the γ coefficient is useful for more than an hypothesis test about timing ability. Since a value of 1 corresponds to the timer effectively providing a whole put on the equity position, any value less implies that the timer is only providing a partial put. Thus, the downward bias in the γ coefficient leads to incorrect inference about the value-added by the manager -- above and beyond the simple question of whether there is timing skill.

The first panel in Table 1 reports the median α values. Notice that the regression using the uncorrected monthly sample appears to attribute the skill to positive alphas. This is consistent with the evidence in Kon (1983), Henriksson (1984) and Jagannathan and Korajczyk (1986) that the timing and selection measures are negatively correlated. Were we to use this test on a mutual fund manager, for example, we might infer that the manager had superior selection ability, as opposed to timing ability. The monthly corrected measure has alphas close to zero, but typically below. This suggests that the corrected test is slightly biased towards finding negative selection ability. This confirms evidence in a number of studies finding negative timing ability.

The panel with β values indicates that market risk exposure for the timer with no skill is effectively the average of the exposures over the interval. When daily returns or corrected monthly
variables are used, the regression appears to successfully distinguish market exposure and timing activity, as the timing ability increases. This is only marginally true when uncorrected monthly returns are used.

III.2 Size and Power

Table 2 focuses on the size and power of tests about $\gamma$ coefficient under the three specifications. The table reports the quantile of the critical t-value of 2 on the timing coefficient in the distribution of t-values generated by 500 simulations for each skill level and for each of the three specifications of the Henriksson-Merton regression. The first column reports results under the null hypothesis of no timing skill. The value of .884 for the “daily” specification indicates that null would have been falsely rejected 21.6% of the time using the traditional 95% confidence level. Thus the power of the daily test is less than the power of the monthly tests, although test size is not held constant to allow direct comparison. The power of the different specifications of the test to detect skill varies dramatically. For example, the null for a timer with perfect foresight is only rejected 40% of the time when the standard Henriksson-Merton specification with monthly data is used. When the corrected monthly data is used, the power increases dramatically. The corrected test never fails to reject the null when the skill is .5 or over. Even when the skill level is .1, the test has some power to reject. The Henriksson-Merton test on daily data is quite powerful, however, he first row of Table 2 suggests that any application of the test should be calibrated appropriately.

IV. Conclusion

Simulations of market timing strategies under reasonable assumptions indicate that the
widely used Henriksson-Merton parametric test has low power to detect timing skill. In fact, for any practical level of skill is has no power at all. In addition, the information about the implied put given by the regression coefficients is strongly biased downwards. The reason for the failure of the statistics is simply that the market timer makes decisions at a more frequent interval than the one over which the value of the implied put is calculated.

We propose a simple correction for the problem which appears to perform quite well. While not as powerful as tests directly on daily timer data, our approach has the advantage of not requiring daily timer data to be calculated. Instead we rely upon an instrument developed from the daily returns to an index correlated to the timer’s risky asset. We find that the HM test with the correction has some power to detect even moderate timing skill.

The logical extension of this paper is to test the market timing skill of mutual fund managers and other investment managers who classify themselves as timers. Before that, however, there are a number of important issues to clarify. Our simulations are based upon a relatively simple model of timers. Few timers are 100% stocks or 100% bills from day to day. We are currently simulating timers who deviate less that 100% conditional upon their forecasts. Preliminary results suggests that this simply weakens the power of all of the tests to detect timing skill.
References


Table 1

The table reports median values of coefficients from Henriksson-Merton non-parametric regressions on simulated data for a range of timing skill. Three specifications of the Henriksson-Merton regression are simulated. The regression equation for “Daily” is \( R_t = \alpha + \beta R_{m,t} + \gamma \max[-R_{m,t},0] + \epsilon_t \) where \( t \) refers to returns calculated daily. “Monthly Sampled” has the same specification, except that the frequency of observation is in monthly. “Monthly Corrected” also uses monthly values for the dependent variable and the market variable, but substitutes the cumulated daily values \( \prod_{\text{month}} \max[1-R_{m,t},1] - 1 \) for the third term in the regression. 500 simulations are performed for each skill level. Ten years of daily market premia are generated as i.i.d. log-normal random variables with annualized mean of .1 and annualized standard deviation of .16. Skill levels range from no foresight “0” to perfect foresight “1”. Intermediate levels are generated by imperfect forecasts as mixtures of the true daily value with a random daily from the ten years sample. For example, “.5” indicates that the timing forecast premium for day \( t \) is \( (.5)R_{m,t} + (.5)R_{m,0} \). We simulate the timer’s strategy as 100% in the risky asset when the forecast premium for the day is positive.

### Alpha Values

<table>
<thead>
<tr>
<th>Skill:</th>
<th>0</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily</td>
<td>0.0000</td>
<td>-0.0003</td>
<td>-0.0011</td>
<td>-0.0012</td>
<td>-0.0006</td>
<td>-0.0001</td>
<td>0.0000</td>
</tr>
<tr>
<td>Monthly Corrected</td>
<td>-0.001</td>
<td>-0.0063</td>
<td>-0.0209</td>
<td>-0.0257</td>
<td>-0.0126</td>
<td>-0.0014</td>
<td>-0.0001</td>
</tr>
<tr>
<td>Monthly Sampled</td>
<td>0.0006</td>
<td>0.0094</td>
<td>0.0312</td>
<td>0.0581</td>
<td>0.0768</td>
<td>0.0837</td>
<td>0.0845</td>
</tr>
</tbody>
</table>

### Beta Values

<table>
<thead>
<tr>
<th>Skill:</th>
<th>0</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily</td>
<td>0.5142</td>
<td>0.6113</td>
<td>0.8409</td>
<td>1.0138</td>
<td>1.0341</td>
<td>1.0045</td>
<td>1.0000</td>
</tr>
<tr>
<td>Monthly Corrected</td>
<td>0.5105</td>
<td>0.6119</td>
<td>0.8601</td>
<td>1.0662</td>
<td>1.1038</td>
<td>1.0787</td>
<td>1.0748</td>
</tr>
<tr>
<td>Monthly Sampled</td>
<td>0.4975</td>
<td>0.5225</td>
<td>0.5917</td>
<td>0.6294</td>
<td>0.6416</td>
<td>0.6382</td>
<td>0.6310</td>
</tr>
</tbody>
</table>

### Gamma Values

<table>
<thead>
<tr>
<th>Skill:</th>
<th>0</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily</td>
<td>0.0000</td>
<td>0.1901</td>
<td>0.6553</td>
<td>1.0132</td>
<td>1.0664</td>
<td>1.0092</td>
<td>1.0000</td>
</tr>
<tr>
<td>Monthly Corrected</td>
<td>0.0041</td>
<td>0.1788</td>
<td>0.6245</td>
<td>0.9935</td>
<td>1.0588</td>
<td>1.0098</td>
<td>1.0018</td>
</tr>
<tr>
<td>Monthly Sampled</td>
<td>-0.0296</td>
<td>-0.0008</td>
<td>0.1247</td>
<td>0.1833</td>
<td>0.1979</td>
<td>0.1914</td>
<td>0.1707</td>
</tr>
</tbody>
</table>
Table 2

The table reports the quantile of the critical t-value of 2 on the timing coefficient in the distribution of t-values generated by 500 simulations for each skill level and for each of the three specifications of the Henriksson-Merton regression described in the previous table. The column under "0" reports results under the null hypothesis of no timing skill. The value of 0.884 for the "daily" specification indicates that null would have been falsely rejected 21.6% of the time using the traditional 95% confidence level. The columns following report the power of the different specifications of the test to detect skill. For example, the null for a timer with perfect foresight is only rejected 40% when the standard Henriksson-Merton specification with monthly data is used.

<table>
<thead>
<tr>
<th>Skill:</th>
<th>0</th>
<th>0.1</th>
<th>0.3</th>
<th>0.5</th>
<th>0.7</th>
<th>0.9</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily</td>
<td>0.884</td>
<td>0.016</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Monthly Corrected</td>
<td>0.982</td>
<td>0.766</td>
<td>0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Monthly Sampled</td>
<td>0.986</td>
<td>0.962</td>
<td>0.878</td>
<td>0.740</td>
<td>0.612</td>
<td>0.556</td>
<td>0.600</td>
</tr>
</tbody>
</table>