Outlier Screening Protocols for Stock Market Studies: A Suggested Screen

Edward J. Lusk[*,1], Michael Halperin[2] and Ivan Petrov[3]

[1]*The State University of New York (SUNY) at Plattsburgh, School of Business and Economics, Plattsburgh, NY, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA*

[2]*Lippincott Library of the Wharton School, the University of Pennsylvania, Philadelphia, PA, 19014, USA*

[3]*Hewlett Packard Inc., Sofia, Bulgaria*

Abstract: In the Data Streaming world, screening for outliers is an often overlooked aspect of the data preparation phase, which is needed to rationalize inferences drawn from the analysis of data. In this paper, we examine the effects of three outlier screens: A *Trimming Window*, *The Box-Plot Screen* and the *Mahalanobis Screen* on the market performance profile of firms traded on the NASDAQ and NYSE. From among seven screening combinations tested, we identify a single screening protocol that is the sequential application of all three screens. This protocol is: (1) simple to program, (2) significantly effective statistically and (3) does not compromise power. This important result demonstrates that for the usual data used by Financial Analysts there is one screening protocol that can be relied upon to satisfy the outlier assumption of the regression model used in generating the usual firm CAPM Return and Risk profile.

JEL: Classification: G11, G12, G32, and G30

Keywords: Beta, Jensen's alpha, SPI, TPI and idiosyncratic risk.

INTRODUCTION

There is a growing literature on the importance of satisfying the data input assumptions of models that are used to generate inference information. One may readily intuit from the following recent research focusing on inference perturbation due to outliers, that if decision makers ignore assumptions that "qualify" the use of inference models, they risk making decisions based on inaccurate or irrelevant information [1-8]. These citations rationalize the logic that modeling assumptions must be satisfied when information is generated that will be used by Financial Analysts [FA] to develop recommendations regarding investment decisions for traded organizations.

In studying a firm's market performance, it is often the case that FA use two general measures of *Return* and *Risk*. *Return* is traditionally calibrated by the following measures of excess return: Jensen's α, the Sharpe Performance Index [SPI], and the Treynor Performance Index [TPI]; for the *Risk* component one usually uses the CAPM β which addresses market-indexed relative risk as surrogated by return volatility. To complete the risk calibration, one uses non-market or non-systematic risk called idiosyncratic risk [iR] which is surrogated by the residuals of the OLS regression [9]. In studies using these Return and Risk measures, the fundamental model used to produce these profiling measures is the OLS one-stage, two-parameter linear regression [hereafter: *β-OLS regression*]. This regression uses, as the dependent variable, the stock market returns of a firm regressed against the independent variable—the time matched returns of the market traditionally surrogated by the S&P500.

In using the β-OLS regression as the information generating system, it is well understood that outliers need to be eliminated as one of the assumptions underlying the β-OLS regression is that the Y_i are independent $N(\mu = \beta_0 + \beta_1 x_1, \sigma^2)$ random variables [10]. The implication of this assumption is that the input data must be outlier-free. This requirement is certainly reasonable in the practical world of the FA because asymmetric outliers, even as few as one, change the character of the β-OLS regression fit by: (1) reorienting the slope of the regression line, (2) affecting the intercept of the fitted line as the slope reorientation will be a ridged-motion rotation, (3) increasing the β-OLS regression variation thus causing the CIs to widen on all estimated parameters and so negatively affecting precision, and finally (4) affecting the nature of the iR— i.e., the residuals of the β-OLS regression filter. Thus the effect of outliers is profound; they affect all of the following Return and Risk measures: Jensen's α, the SPI, the TPI, the regression slope—β, and iR.

This is the point of departure of our study, which addresses screening for outliers when the β-OLS regression is used by FA as the inference model for market studies of traded firms. Specifically, the purpose of our study is to:

1. Test the often used outlier screening protocols: *A Trimming Window* [TW], *The Box Plot* [BP], and *The Mahalanobis D-Measure Screen* [MS] to determine if, in screening for outliers, the screening protocol that satisfies the outlier-assumption of the β-OLS regression changes depending on the particular

*Address correspondence to this author at the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, USA; Tel: +01.518.564.4190; Fax: +01.518.564.4215;
E-mail: lusk@wharton.upenn.edu

1874-9151/11 2011 Bentham Open

Return or Risk measure: Jα, SPI, TPI, β or iR under consideration. We label this as *the efficiency dimension*. We are using the term *efficiency* as it relates to the data streaming world where more or less continuous data is streamed as EDT from various sources, such as *Bloomberg*™, that are often subscribed to by the FA community. We label a single screen that is found to satisfy the outlier-assumption independent of the particular Return or Risk measure as the most efficient screen. Efficiency calibration is critical because in a data streaming world the FA cannot spend valuable decision making time deciding upon which screen to use in the data preparation step. The reason for this is that ever since automatic trading algorithms have been in use, circa 1980—i.e., the expert-system-algorithmic versions of the "day-trader"—the time-to-decision has become very short. See [11] for an excellent discussion of the time-to-decision issues in the trading milieu. Therefore, for our purpose efficiency will be calibrated based upon the number of different outlier screening alternatives that are required in using the β-OLS regression as the information generating model. The highest level of efficiency then will be if one screen is unconditionally recommended as the necessary screen over all five of the performance measures tested,

2. Determine if the parameters of the recommended screen(s) from the above set of possible screens could be programmed without decision making intervention from the FA or extensive search conditional-analyses. We label this as the *simplicity dimension*. The simplicity dimension is then the possibility of creating an Expert System's Screening Protocol whose parameters can be automatically set by the inputted data stream and so can be sensitive to the time-to-decision-imperative. The simplicity issue is fundamental in a data streaming world. Because time is critical, screens that cannot be programmed, and therefore require unique interventions by the FA to set screening parameters, will not be useful,

3. Determine if there is a statistically significant inferential impact of the screen(s) identified as efficient and simple on the Return and Risk measures. If it is determined that there is no inferential impact between the screened data and the unscreened data—the benchmark—then the screening issue is moot. We label this as the *effectiveness dimension*, and

4. Determine if there are statistical power differentials in the application of the various screens—*the power dimension.* Here we are interested in detection power relative to the Return and Risk measures as we will be electing outlier elimination rather than Winsorizing-replacement. In this context, it is well understood that such screening selectively affects variance. There is a power trade off—as the screen removes data points power falls; however, at the same time as the screen targets high variance data points power is increased. Depending of the results of our screening analysis regarding efficiency, simplicity, and effectiveness there may be power issues to be considered.

This information will then be summarized to answer the questions: Does screening matter? And if so, what is the screening configuration revealed by the analysis; and what are its characteristics regarding efficiency, simplicity and power?

To be clear as to the nature of this study, we wish to emphasize that our focus addresses outliers in market studies using the β-OLS regression to form information on the Return and Risk measures: Jensen's α, the SPI, the TPI, β, and iR. This focus limits our investigation in the following important way: We cannot speak to how our β-OLS regression outlier screening results would compare to: (1) different modeling systems that may be used to measure the five Return and Risk measures, nor (2) other measures of Return and Risk.

Consider now an "*en bref*" for the following sections of the paper. In the section **Elucidation of the Screening Measures**, we detail the outlier screening protocols that will be used in filtering the data used in the β-OLS regression. We will present how these screening filters are operationalized as the focus of the study is to provide guidance useful in parameterizing these screens. In the next section: **Return and Risk Measures**, we describe the nature of the classic Return and Risk measures often used by the FA as the decision making variables underlying their stock recommendations. A central question of our study is: Are these Return and Risk measures affected by the screening required by the β-OLS regression? If not, the question of screening is moot. In this regard, we will determine: (1) what is the "final" screen that is needed, and (2) are there statistically significant differences for the Return and Risk measures between the No Screening alternative and the Final required screen. Continuing in the next section: **The Study: Validation Hypotheses and Results**, we describe the dataset used in the study. We have selected as a "validation" time frame a five-year period in the exuberant market of the Bubble-Build-Up in the USA. An operational validation of the screens is the fact that they behave as expected for the data in this accrual period.

Because this time frame is limited to five years, we have used daily market returns to increase detection power. Further, daily return data is the least smoothed return activity available from our EDT download source: WRDS™ which also has monthly return data. Using relatively un-smoothed returns aids comparatively in measuring underlying time-related volatility association. Finally, daily data is used by most of the data reporting services which report measures of β—for example, CRSP™, Ibbotson™, and Morningstar™ to mention a few.

In what follows, we present evidence for the single recommended screen which we find to be the sequential application of the Trimming Window, the Box-Plot and the Mahalanobis Screen: TW&BP&MS. This combination, which comparatively screens the most data points, is: (1) the most efficient, in that it is the same for each of the five performance measures, (2) simple, in that all the sequential filtering blocks can be programmed in Excel, (3) effective as there are meaningful statistically significant differences between the Screening and Non-Screening results, and (4) not found to compromise power. Finally in the section: **Summary and Conjecture for the Future: Suggested**

Extensions, we close the loop by summarizing the study and offering suggestions for further investigations.

ELUCIDATION OF THE SCREENING MEASURES

We now present the three filters used in our study.

Winsorizing and Trimming

Winsorizing was first introduced by [12] who notes:

> "Winsor and perhaps others have suggested using for the magnitude of an extreme, poorly known, or unknown observation the magnitude of the next largest (or smallest) observation. We shall show that when symmetry is maintained (or proper adjustment is made) this practice results in estimators of the mean whose efficiencies are scarcely distinguishable from those of best linear estimators. For non-symmetrical censoring, it is demonstrated that optimum simple estimators of the mean result from these 'Winsorized' estimators".

One recognizes that Winsorizing is just the trimmed mean transformation introduced by [13] or spectral windowing often used in frequency or periodogram studies [14, 15] except *that in Winsorizing there is a data specific replacement required for the windowed data points*. This creates a slight problem if FA are trying to be sensitive to the moral hazard issue of agenda-serving selective data replacement. This moral hazard possibility may be simply avoided in correlation based studies, such as estimating CAPM market performance, as the strict time series requirements do not have to be served as CAPM inferences are drawn from the β-OLS regression model which is founded on correlation and not on time series equal-spacing. Therefore window-trimming [WT] is preferred over Winsorizing in our study context as it is more objective.

As a final note on replacement and elimination; both usually increase precision because the standard error reduction from the variance eliminated is a direct effect for Winsorizing and for Trimming is almost always greater than the standard error increase that occurs due to the reduction of the sample size. This tradeoff in the Trimming case of course raises relative power as an issue which we will explore. For example, according to [3]:

> "To examine the effects of outliers in the asset growth distribution, we winsorize the asset growth distribution at the 1% and 99% points of the distribution. Winsorizing the data has the effect of making the asset growth relationship stronger." p. 1632

Consider now our trimming window.

The Trimming Window

The Empirical Rule [ER] introduced by Abraham De Moivre (1667-1754) [16], simply states that very often the distribution of collected observations may be characterized using the Mean and the Standard Deviation [Sd] as follows:

68% of the observed data fall into the interval: [Mean ± 1Sd]

95% of the observed data fall into the interval: [Mean ± 2Sds]

99% of the observed data fall into the interval: [Mean ± 3Sds]

Using this remarkable empirical observation, an Empirical Rule Trimming Window [ERTW] for our study was formed as follows: Assuming that a reasonable parameterization of the ERTW [TW hereafter] is to conserve 95% of the data, DS(t), then a TW that screens for 2½% of the data on the high as well as on the low side is a reasonable place to start as the first filter. The TW-parameterization for all time points, t, is:

IF [DS(t)] > [Mean + 2 x Sd], then eliminate DS(t);

IF [DS(t)] < [Mean – 2 x Sd], then eliminate DS(t);

IF Not then use DS(t) in the modified DS.

One recognizes that the TW is a *parametric screen* as it is calibrated using the mean and standard deviation of the data in the data stream.

The next outlier filtering steps recommended are to identify relative distance outliers in (1) the histogram placement in Cartesian Space—i.e., one dimensional and so a non-relational measure and (2) relational or specifically Correlation Space. These two screens provide conceptually independent measures of possible outliers compared to the TW which is a non-relational parametric screen.

The Box Plot

The Box-Plot [BP] was developed by John Tukey [1915-2000]; they are called Box-Plots because they are shaped like a box. The BP is a median centered measure that uses a fixed expansion of the Inter-Quartile Range [IQR] to construct an interval outside of which the points are identified as BP-outliers. We have selected this measure because the TW is parametric as it uses the arithmetic mean and standard deviation to create the trimming interval whereas the BP uses the median and the IQR as the location and dispersion metric and so gives a "*Non-Parametric*" perspective. This adds a robustness dimension to the data modification procedures.

The BP is in the SAS/JMP™ *Data Description* APP-Platform and also is easily programmed in Excel [the MS-Office™ suite]. We will be using the SAS/JMP™-default parameterization that sets the outlier screening interval at:

BP Window: [(25^{th} Percentile Data Pt – 1.5 x IQR) to (75^{th} Percentile Data Pt + 1.5 x IQR)]

Any D(t) value that falls outside of this interval is marked as a BP outlier. These limits are sometimes called "Whisker-Limits."

The Mahalanobis Screen

The Mahalanobis Screen [MS] is a Correlation Screen due to Prasanta Chandra Mahalanobis (1893-1972) [17]. The MS screens outliers in the Pearson product moment space—or what are called "correlation outliers"; as such the MS is formed on a different basis than the TW or the BP. We recommend the Mahalanobis Screen [MS] as it is part of the SAS/JMP™ APPs. The programming for the MS may be found in [18] and will thus aid respecting the simplicity dimension. As the MS is the third filter, it is not likely to be affected by extreme outliers. Our testing shows that the MS

set at the 95% confident level and the 95% Pearson Probability Ellipses are essentially the same in the percentage detection of outliers.

Consider now the definition of the Return and Risk measures that are the standard fare for the FA in profiling the performance of firms traded in active markets.

RETURN AND RISK MEASURES

Excess Return Measures

Jensen's alpha, noted as Jα, is a measure of the market benchmarked excess return over the projected CAPM return. There are two ways to create Jα. When one has a time-matched measure of the risk-free rate, such as the 30-day US Treasury Bill rate, one can create excess returns by subtracting the time matched risk-free rate from the firm and from the market return time series, and then run the usual β-OLS regression; the intercept will be the measure of Jα. Another way, that we used, is to use the following formula due to [19].

EQ1 $J\alpha = \bar{r}_f - (\bar{r}_{rf} + \hat{\beta}_f[\bar{r}_m - \bar{r}_{rf}])$

where:

$\bar{r}_f$ *is the average return of the firm,*

$\bar{r}_{rf}$ *is the average risk − free return,*

$\hat{\beta}_f$ *is the estimated CAPM firm beta/slope,*

$\bar{r}_m$ *is the average return of the market.*

This form best indicates the nature of Jα as the excess of the average return of the firm, $\bar{r}_f$, over the projected average return, $\bar{r}_{rf} + \hat{\beta}_f[\bar{r}_m - \bar{r}_{rf}]$. So Jα gives an indication of the return performance of the firm relative to the return of the market portfolio after considering the risk-free rate. A positive (*negative*) Jα indicates that the firm outperformed *(was outperformed by*) the market portfolio projection respecting excess return.

The Sharpe Performance Index (SPI) is measured as:

EQ2 $SPI = [\bar{r}_f - \bar{r}_{rf}]/s_f$

where:

s_f *is the standard deviation of the returns of the firm*

and the **Treynor Performance Index** (TPI) is measured as:

EQ3 $TPI = [\bar{r}_f - \bar{r}_{rf}]/\hat{\beta}_f$

These indices give the firm's average return over the average risk-free rate, noted as excess return, as a percentage of firm risk. The SPI is the excess return of the firm relative to its total risk, as measured by the standard deviation of the returns of the firm. The standard deviation of returns is the usual definition of total risk due to [20]. Thus, the SPI is a measure of excess return as a percentage of *total* firm risk. The TPI uses the same numerator as does the SPI, but divides it by the firm's period beta, which is, as discussed above, the index multiplier of the relative return of the firm compared to that of the market portfolio. In this sense, the TPI is excess return as a percentage of non-diversifiable risk or systematic risk, whereas the SPI is indexed on total risk—excess return relative to total firm risk. These are the standard performance-index comparisons that have been used for more than 45 years to judge the relative performance of organizations as calibrated on volatility or risk of the firm.

Risk Measures

The Capital Asset Pricing Model (CAPM) **Beta**, β, is the slope of the OLS regression of the matched firm time series with the market where one uses for the market the S&P500. It is the classic measure proposed by Sharpe, and, aside from its use in the development of the CAPM as a theoretical construct as it relates to the EMH, β is simply a ratio measure of co-variation relative to market variation where variation is the Markowitz surrogate for risk. In this way, when β = 1 then the firm and the market have the same risk, when β > 1 then the firm is riskier than is the market, and when β < 1 the firm is less risky than the market. We will use β in this relative risk/volatility sense which is a more general conceptualization than β as the central CAPM feature where β is used to generate the return projection: $\bar{r}_{rf} + \hat{\beta}_f[\bar{r}_m - \bar{r}_{rf}]$. Simply said, β is the slope of the β-OLS regression, and is therefore the relational multiplier between the returns of the Firm and the Market-matched time series.

Our measure of **Idiosyncratic Risk** is offered by [21]. Computationally it is:

EQ4 $iR_{B\text{-}H/L} = s_f - \hat{\beta}_f \times s_m$

where: s_f and s_m are the standard deviations of the returns of firm and the market, and $\hat{\beta}_f$ is the slope of the OLS-regression filter for the firm and the market returns.

This Ben-Horim and Levy measure is preferred to the classic measure of iR—the Sharpe measure—as this latter measure has been shown to be biased on the high side [21, pp. 293-4].

THE STUDY: VALIDATION HYPOTHESES AND RESULTS

The Accrual of Study Firms

Because our interest is in the effect of screens on the Return and Risk measures, we have selected a time in the market where we have an *a-priori* sense of market direction, and can therefore anticipate the effect of screening on the financial profiling measures. This will be important as an operational validity check on our results. For this reason we have selected the Internet-Bubble-Build-Up Period [hereafter *the Bubble Period*] from 1 January 1995 to 31 December 1999. This was a time when stocks were experiencing extraordinary growth—albeit unrealistic—and so we will have an *a-prior* expectation that screening for outliers will produce a mollification of Excess Returns as well as Relative Risk. Simply said: The three screens: TW, BP and MS should differentially screen high-side outliers in the Bubble Period and so reduce *Excess Returns* and attenuate Volatility and so reduce *Relative Risk*. We will use this *a-priori* expectation to validate the results that we generate using the three screens over the five market performance measures.

To produce screening results that are generalizable, we have selected three grouping of firms: The Old Economy [OE], IPOs opening during the accrual period, and firms from the New Economy [NE]. For accruing these firms, we followed guidelines and suggestions found in [22-24]. OE firms were in the durable goods sectors such as Metals, Heavy Manufacturing, Mining and Chemicals. For the New Economy, we used firms that are in the Technology, Electronics and related Light Manufacturing, Software, as well as Systems Development sectors. In addition, we added a set of non-dot.com technology-related IPOs that opened in 1993 and traded on the NYSE or the NASDAQ during the Bubble Period from 1 January 1995 to 31 December 1999.

We found that there were 30 firms that fit the usual profile of the OE that were in operation in 1980 and that were continually listed on the NASDAQ or NYSE until the end of 1999—i.e., firms that exhibited long-term stability. For the NE, we found that there were 34 firms that fit the Standard Industrial Code [SIC] rubric for the NE; finally there were 34 IPO-firms opening during 1993 that survived from 1 Jan 1995 until 31 December 1999. This produced a dataset of 98 firms measured on five market performance measures using seven screens or 3430 values [98×5×7]; this dataset is included on Scholarly Commons: http://repository.upenn.edu/ and is available as an unrestricted download where we waive any intellectual property rights.

Our Validation Hypotheses

There are several screening protocols that could be used. The screens could be individual i.e., only one screen, or applied in a sequence—i.e., in pairs or successively. For the three screens that we will be using, there are 15 such possibilities where order is considered. We have selected the following seven outlier-testing protocols:

No Screening,

TW, BP, or MS applied individually

TW&BP, the TW&MS

TW&BP&MS.

In all cases, we have begun with the TW, which should create the widest or most liberal window, as it is variance driven, and outliers increase variance; we have ended with the MS screen as it is sensitive to extreme outliers, and therefore works more effectively as the last screen in a sequence. These selection conditions reduce the 15 screens to the six noted above. Also, there is only one screening application using the TW. Because the TW is set on the standard deviation of the series, re-screening using the TW would eventually "eliminate" most of the data. This is why trimmed mean applications are usually restricted to a first-pass-only.

Validation Information

As indicated above, we have selected the Bubble Period to research the effect of the screening protocols. Therefore, let us first present the validation information—i.e., the screens work as expected in the accrual period. Specifically, given the accrual period, we expect that the more points that are screened the lower will be the values of the Return and the Risk measures—that is to say, the value-slope will be negative with respect to the number of points removed.

For all of the Return and the Risk measures (with the exception of the TPI) the slope of the regression for the measured value relative to the number of data points removed by screening was in the expected negative direction, and was statistically significant at $p < .1$.

For example, consider the Plot of Jα:

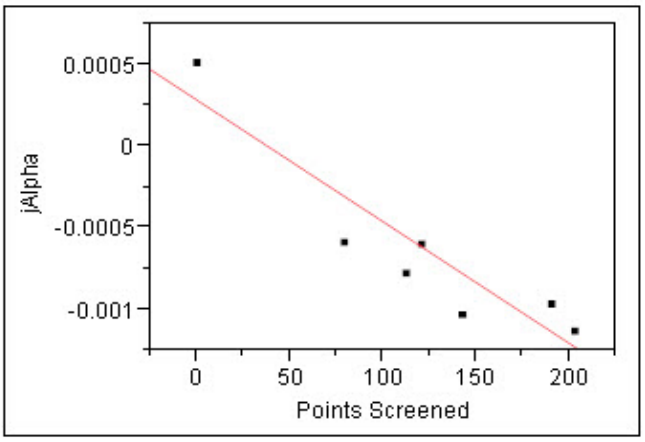

Fig. (1). Jensen's alpha by points screened.

For this regression, the adjusted R^2 was 0.82 and the p-value for the negative slope was 0.002, indicating that there is strong evidence that the more points that are screened the lower will be the estimate of excess return as measured by Jensen's α. This confirms our *a priori* expectation for the accrual period, and is a positive affirmation/validation of the directional sensitivity of the various screens.

To determine the overall relationship between the five performance measures and the seven screening protocols, we first determined for each measure the number of points screened, and then ranked them from lowest to highest. This information is presented in Table **1**.

Table 1. Median and Mean Number of Points Used Over the Seven Screening Protocols Ranked from Highest to Lowest for all 98 Firms Over All Five Return and Risk Measures

Measure	Median Points	Mean Points
No Screening	1,263	1,263
MS	1,184	1,184
BP	1,150	1,148
WT	1,142	1,142
WT&BP	1,121	1,117
WT&MS	1,072	1,062
WT&BP&MS	1,060	1,055

Consistent with the information in Fig. (**1**), Table **1** also provides confirmatory validation information in that:

1. as expected, compared to *No Screening*, applying all three screens in series resulted in the fewest number of points for analysis.

2. because the Mahalanobis Screen [MS] is correlational in nature, and so requires two outliers in combination in the correlation space, the MS screens removes the minimal number of points relative to the *No Screening* option,

3. the individual Screens: TW, BP and MS screen fewer points than the application of any of the ordered combinations: TW&BP, TW&MS and TW&BP&MS.

This information validates the sample as a typical characterization of the Bubble Period in that the screens follow their operational logic regarding the number of point differentially removed, and therefore provide the following context for the application of the various screens: The more points that are screened, the lower will be the Return and Risk profile. If this were not the case, it would call into question the operational functioning of the screens, and thus call into question the logic of using screens that seemed to be directionally insensitive to the five Return and Risk measures.

The critically important questions that we will now address based on this operationally validated dataset and screens are: (1) At what point in the application of these seven screens do we find that no outliers remain? (2) At what point, are the performance measures statistically significantly different compared to the No Screening protocol? and (3) Are there power differentials over the various screens? Consider these questions now. This information will be essential in rationalizing a Screening Rule that can be recommended to the FAs conducting market studies.

PRINCIPAL RESEARCH QUESTION AND RESULTS

The first question is, at what point do we stop detecting outliers? As we know from Table **1**, on average each of these screening protocols "progressively" removes a certain number of points, and this continues until the application of the final screen: WT&BP&MS. To further investigate this question, we examined over all the Return and Risk performance measures how often, after applying the WT and then the BP screen, that there were additional outlier points identified by the MS screen. Remarkably 100% of the time the MS screen identified outliers after the application of the WT followed by the BP. Therefore the first question has a clear answer. For our sample, WT&BP&MS always screens additional points compared to the WT&BP screen. This suggests that the final screen, WT&BP&MS, is recommended in order to rationalize or justify the use of the β-OLS regression. We label the screen: WT&BP&MS the ***necessary screen***. Further, as there is only one screen, it is the **most efficient**. Because it can be programmed to be parameterized by the inputted data, it is also **simple**.

The next question is the effectiveness issue: Are the five performance measures statistically significantly different as between the NO Screening alternative and the necessary screen. To answer this question see Table **2** following, where the median values under the two different screening protocols are contrasted using the conservative Wilcoxon Rank Sum test. The p-values reported are two-tailed.

Table 2. Median Test of the Extremes: No Screening Compared to WT&BP&MS

Performance Measure	No Screening Median [IQR]	WT&BP&MS Median [IQR]	P-Values
Jα	0.0005 [0.0013]	-0.001[0.002]	<0.0001
SPI	0.03[0.034]	-0.02[0.082]	<0.0001
TPI	0.001[0.0018]	-.001[0.0044]	<0.0001
Beta	0.68[0.587]	0.55[0.487]	0.025
iRisk	0.033[0.0249]	0.022[0.0163]	<0.0001

*Discussion of the Screening results in Table **2**.* The results of the statistical comparison between the No Screening protocol and the consecutive application of the WT, then the BP and finally the MS, i.e., the necessary screen WT&BP&MS, is dramatic and conclusive. In all five cases, there is a statistically significant result in the anticipated direction between the two screening protocols and therefore the screen is effective. The implication for the FA is that screening is needed to satisfy the β-OLS regression requirements, and that this necessary screen will have a dramatic effect on the market performance profile of the firm. In so doing it will better serve the FA in developing investment recommendation strategies.

In summary, the WT&BP&MS screen removes the most points, on average 16.1%. In addition, the screen removes points at each screen application. Finally, the measured values of the Return and the Risk for the accrued firms are highly statistically different compared to the No Screening protocol. We offer that *to best satisfy the β-OLS regression requirement of drawing inference from data that is relatively outlier free,*

> ***the FA should use the WT&BP&MS screen to measure the market profile performance statistics of Excess Return: Jα, SPI and TPI and Risk: β and iRisk.***

POWER INFORMATION

To finish our investigation, we have also investigated the relative power of the various screens. The test here is to determine if the detection power differs in an important way for the various screens. As introduced above, we know *a-priori* that the Confidence Interval on the estimated parameters will shrink where the screens remove outliers because, in any practical case, the range will decrease and consequently the variation will decrease at a rate which will outpace the loss of power due to the reduction of the sample size.

Therefore, as a final piece of information, we are interested in determining if there is a detection or power effect between the three Single screens and the three Multiple screens. This is a conceptual robustness "What-if-Analysis" regarding power. We have already determined that there is one effective and simple screen: WT&BP&MS. We now ask: If one had the choice of any of the screens, would there be power considerations in the selection?

To determine what actually happened in the power context for the five measured market performance measures relative to the six screens, we used both the differential effects on the IQR and the parametric Precision for over all the firms accrued. To be clear, we have six screening protocols; here we exclude the Non-Screening protocol, as we are interested in the screening effects given the results reported in Table **1**. Each of the six screening protocols develops a CI and has an IQR for the 98 firms for each of the five Return and Risk measures used in the study or 30 [6 x 5] CI and IQR measures for each firm. In Table **4** the results are reported for the **Single** screens [WT, BP and MS] contrasted with the **Multiple** screens [WT&BP, WT&MS and WT&BP&MS].

Table 3. Single *vs* Multiple Screens: Relative Effective Power

	Single[TW,BP,MS]:Multiple[TW&BP, TW&MS, TW&BP&MS]	P-Value
IQR	11.8 : 11.7	0.996
CI Width	3.4 : 3.8	0.846

*Table **3** Discussion.* The two-tailed p-values make clear that there is no evidence that there is a power difference between the two screening groups tested. Both the IQR and the Width of the CIs are essentially the same for the two screening groupings: Single and Multiple Screens. For example, this result indicates that the average IQR over all the five market measures for all the firms tested is around 12 IQR units for both screen groups. The implication of this result is that there is no advantage in searching for screens interior to the final recommend screen: WT&BP&MS. This reinforces the conclusion of the study, namely: that the WT&BP&MS combination screens are not only simple to apply, efficient and effective, but also do not compromise power.

SUMMARY AND CONJECTURE FOR THE FUTURE: SUGGESTED EXTENSIONS

We began this study to determine if there was a particular outlier screen that FA could apply in conducting market studies using the β-OLS regression as the inference generating model. We examined three screens:

1. A Trimming Window the width of which was calibrated using the 95% Empirical Rule of the Mean ± 2 x [the standard deviation of the un-adjusted series]; this is a ***parametric window***,
2. An outlier screen founded on the Box Plot which uses the Median and the IQR, and so is ***non-parametric*** in nature, and finally
3. A Relational screen due to Mahalanobis which uses ***relational*** outliers from the two data sets.

The results are simply summarized as:

> **For the various screening protocols tested, the three-staged sequential screen: WT&BP&MS is efficient, simple, effective, and does not compromise power. This is the screen recommended independent of the Return and Risk market measures tested.**

Discussion and Extension. Is this principal result the end of the screening investigation? The answer to this rhetorical question is, of course, No! We view this research report as the beginning of the collection of information that we hope will lead to Meta-Analyses addressing data preparation that will help the FA to better navigate in the data-streaming world where the modeling system of choice is the β-OLS regression. We hope that other researchers will investigate not only other Return and Risk measures, but of course other time-event periods. In this latter regard, we recommend the post-Sarbanes-Oxley event space: 2003 to just before September 2008 when the *Lehman Bros, LLP* sub-prime debacle almost crashed the financial world; its effects are even being felt today! Another important event time period is the "Post-Sub-Prime-Debacle" that starts with the shocking revelations of massive defalcations and fraudulent reporting at the country level. One marks this at around 2009 to date; we are enamored with the *Economist's* acronym: the PIIGS Block: Portugal, Ireland, Italy, Greece and Spain whose actions may call into question the viability of the EURO.

Finally, thinking outside the "β-OLS regression box", there is the possibility of changing the analytic information generation framework [25] by putting into play a two factor modeling system. If one can "re-invent" an asset pricing model around a regression that is not as sensitive to outliers perhaps with a different loss function, then possibly outlier screening issues can be rendered moot; this is the direction that effectively the Ben-Horim/Levy re-calibration of idiosyncratic risk offers. Moving away from the classic Markowitz definition of risk and moving to perhaps, the Knightian view of Uncertainty, will indeed be difficult; the β-OLS regression inertia is pervasive and seems to be as close to the "immovable object" as any construct in finance. But perhaps we are at a point of departure from the market world envisaged by Sharpe [26], Lintner [27], and Mossin [28].

To this end, we find Regression Ranking Models a viable and interesting "path-less-traveled". In this regard, we are enthusiastic about the work of [29] who note:

> "We suggest the use of ranking-based evaluation measures for regression models, as a complement to the commonly used residual-based evaluation. We argue that in some cases, such as the case study we present, ranking can be the main underlying goal in building a regression model, and ranking performance is the correct evaluation metric. However, even when ranking is not the contextually correct performance metric, the measures we explore still have significant advantages: They are robust against extreme outliers in the evaluation set; and they are interpretable."

They continue referencing the work of [30]:

> "A commonly used definition of robustness in model fitting uses the concept of fitting breakdown point. The breakdown point of a fitting procedure is the % of data points that

> must be arbitrarily badly corrupted before the fitted model is arbitrarily badly corrupted.
>
> Using this standard definition they continue: "linear regression with squared error loss has a breakdown point of 1/*n*. Thus, this is a non-robust procedure—one corrupted data point can affect the fitted model arbitrarily. Linear regression with absolute loss, on the other hand, has a breakdown point of "almost" 1/2. This is a robust fitting procedure, since as long as less than half of the data points are corrupted, we are guaranteed to remain 'reasonably close' to the uncorrupted solution."

Therefore, this sort of "loss-function-liberated" regression model may be a productive start; in the meantime, for the practical needs of the FA working within the β-OLS regression box in the streaming world outlier ***Screening*** appears to be the only practical option.

ACKNOWLEDGEMENTS

We wish to thank Professors Neuhauser of the Department of Finance, Lamar University, Beaumont Texas, USA and Lee, Chuo-Hsuan, The Department of Accounting SUNY: Plattsburgh, Plattsburgh, NY, USA, Mr. Frank Heilig, M.Sc. Risk Controller: Risk Management Section: *Stadtsparkasse GmbH*, Magdeburg, Germany, Ms. Chen, Li Financial Analyst: *Seneca College*, Toronto CA, and two anonymous reviewers for their detailed and constructive comments.

REFERENCES

[1] Fama E, French K. Dissecting anomalies. J Financ 2008; 63: 1653-78.

[2] McInnis J. Earnings smoothness, average returns, and implied cost of equity capital. Acc Rev 2010; 85: 315-41.

[3] Cooper M, Gulen H, Schill M. Asset growth and the cross-section of stock returns. J Financ 2008; 63:1609-51.

[4] Campbell Y, Hilscher J, Szilagyi J. In search of distress risk. J Financ 2008; 63: 2899-939.

[5] Cowan A, Sergeant A. Interacting biases, non-normal return distributions and the performance of tests for long-horizon event studies. J Bank Financ 2001; 25: 741-65.

[6] Lusk E, Halperin M, Heilig F. Market and financial performance as related to idiosyncratic risk and the effect of outlier screening in market studies. J Financ Manage Anal 2009; 22: 59-69.

[7] Dlugosz J, Fahlenbrach R, Gompers P, Metrick A. Large blocks of stock: Prevalence, size, and measurement. J Corp Financ 2006; 12: 594-618.

[8] Filzmoser P, Garrett R, Reimann C. Multivariate outlier detection in exploration geochemistry. Comp Geosci 2005; 31: 579-87.

[9] Lusk E, Halperin M, Bern M. Towards reformulation of the capital asset pricing model (CAPM) focusing on idiosyncratic risk and Roll's meta-analysis: A methodological approach. J Financ Manage Anal 2008; 21: 1-23.

[10] Tamhane A, Dunlop D. Statistics and data analysis. Upper Saddle River, NJ: USA Prentice Hall 2000.

[11] Clark C. Controlling risk in a lighting speed trading environment. Federal Resrve Bank China 2010; 275: 1-4.

[12] Dixon W. Simplified estimation from censored normal samples. Ann Math Stat 1960; 31: 385-91.

[13] Blackman R, Tukey J. Particular pairs of windows in: The measurement of power spectra, from the point of view of communication engineering. Dover, NJ. USA: Dover Press 1958.

[14] Bloomfield P. Fourier analysis of time series: An introduction. New York, NY: John Wiley & Sons 1976.

[15] Jenkins G, Watts D. Spectral analysis and its applications. New York, NY: Holden-Day 1968.

[16] Hald A. History of mathematical statistics from 1750 to 1930. New York, NY: Wiley and Sons 1998.

[17] Mahalanobis P. On the generalised distance in statistics. Proceedings of the National Institute of Sciences of India 1936. http://ir.isical.ac.in/dspace/handle/1/1268.

[18] Sall J, Creighton L, Lehman A. JMP Start statistics. Pacific Grove, CA. Thomson: Brooks/Cole, The SAS Institute 2005

[19] Nielsen L, Vassalou M. Sharpe ratios and alphas in continuous time. J Financ Quant Anal 2004; 39: 103-15.

[20] Markowitz H. Portfolio selection. J Financ 1952; 7: 77-91.

[21] Ben-Horim M, Levy H. Total risk, diversifiable risk and nondiversifiable risk: A pedagogic note. J Financ Quant Anal 1980; 15: 289-97.

[22] Chen, Y-R. Corporate governance and cash holdings: Listed new economy versus old economy firms. Corp Gov: Int Rev 2008; 16: 430-42.

[23] Zanini M, Lusk E, Wolff B. Confiança dentro das Organizações da Nova Economia: uma Análise Empírica sobre as Conseqüências da Incerteza Institucional: Trust within the Organizations of the New Economy: an Empirical Analysis of the Consequences of Institutional Uncertainty: RAC 2009; 13: 72-91.

[24] Jungqvist A, Wilhelm W. IPO pricing in the dot-com bubble. J Financ 2003; 58: 732-52.

[25] Fama E, French K. Common risk factors in the returns on stocks and bonds. J Financ 1993; 33: 3-56.

[26] Sharpe W. Capital asset prices: A theory of market equilibrium under conditions of risk. J Financ 1964; 19: 425-42.

[27] Lintner J. The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets. R Econ Stat 1965; 47: 13-37.

[28] Mossin, J. Equilibrium in a capital asset market. Economics 1966; 34: 768-83.

[29] Rosset S, Perlich C, Zadrozny B. Ranking-based evaluation of regression models. Knowl Inf Syst 2007; 12: 331-53.

[30] Hampel F, Ronchetti E, Rousseeuw P, Stahel W. Robust statistics: the approach based on influence functions. New York, NY USA: Wiley 1986.

Received: November 1, 2010 Revised: February 25, 2011 Accepted: February 28, 2011