Applied Financial Economics, 2008, **18**, 1739–1753

Routledge
Taylor & Francis Group

Deflator selection and generalized linear modelling in market-based regression analyses

Changbao Wu[a,*] and Bixia Xu[b]

[a]*Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada*
[b]*School of Business and Economics, Wilfrid Laurier University, Ontario, Canada*

The scale factor refers to an unknown size variable which affects some or all observed variables in a multiplicative fashion. The scale effect studied by several researchers in market-based regression analyses is defined here as the intriguing combination of coefficient bias and heteroscedasticity caused by the scale. Deflation is the most popular technique used in previous market-based studies to mitigate the scale effect. Selection of a suitable deflator, however, remains as a difficult and challenging task due to the lack of a general statistical framework for this type of research. In this article, we establish a general statistical framework for deflator and model selection. We argue and show that the existence and severity of the scale effect can be identified and measured using the *Average Absolute Values of Studentized Residuals* and the *Relative Total Prediction Error* for stratified firm groups. The proposed framework consists of five major components. Results from our simulation studies and sensitivity analyses show that if the true scale variable is used as a deflator to produce one of the deflated candidate models, this model can be correctly identified using the proposed strategy, even if the working model is mildly misspecified. In addition, our studies show that the generalized linear modelling method can be very useful for mitigating the scale effect when the unknown true scale variable is related to the whole set of independent variables through the so-called mean function.

I. Introduction

In market-based regression analyses, sample data are often cross sectional, with information from various balance sheets and income statements of firms with different sizes. One of the important econometrical issues is to identify and mitigate the so-called 'scale effect' which, if exists, can cause coefficient bias and model inefficiency (Barth and Kallapur, 1996).

Deflation is the most popular method used in this type of research to mitigate the 'scale effect'. In the special case where the scale factor is known, deflation simultaneously cures coefficient bias and heteroscedasticity and is unambiguously the better remedy (Barth and Kallapur, 1996). However, the scale factor is rarely known in reality. One needs to search for a scale proxy and uses it to deflate the original model. Examples of deflators used in prior market-based

*Corresponding author. E-mail: cbwu@uwaterloo.ca

studies include market value of equity, book value of equity, earnings, cash flows, sales, total assets and number of shares outstanding. Christie (1987) was one of the early studies which raise the concern that inappropriate deflator can cause spurious inference (this concern is confirmed by results reported in Sections III and IV of the present study). Since then, a few studies have made efforts to explore the best deflator among alternatives primarily from the economic point of view. The empirical evidence and resulting conclusions are often controversial and mixed. Easton (1998) suggests that closing book value is a suitable deflator. Barth and Clinch (2001) using simulated data suggest that Easton does not demonstrate that deflating by book value produces superior results. Lo and Lys (2000) argue that opening market value is the best deflator and also that deflating by opening market value produces a theoretically more appealing coefficient for dividends in a regression of market value on book value, earnings, dividends and capital contributions. In contrast, Easton and Sommers (2003) argue and find supportive evidence based on the US data that the scale and then the best deflator is end-of-period market capitalization. They strengthen their findings by providing evidence on the superiority of using market capitalization as the deflator over using book value and earnings. However, in the discussion of Easton and Sommer's work, Akbar and Stark (2003) show that end-of-period market capitalization fails to outperform book value and earnings when the UK data are used, suggesting that Easton and Sommer's findings may not be generalized into other countries, which, in fact, questions Easton and Sommer's message that end-of-period market capitalization is the true scale factor.

The controversial discussions and mixed evidence call for further development of a unified theory for deflator selection. Given the fact that the true scale is unobservable, the key practical issue lies in the selection of an appropriate scale proxy (i.e. deflator) among alternatives that can best mitigate the scale effect and reduce the likelihood of invalid and spurious inference. In this study, unlike previous studies which discuss individual deflators in a one-at-a-time manner, we establish a general statistical framework aimed at providing a systematic and operational guidance for deflator and model selection[1] to mitigate the scale effect under any economic situations. We define the scale effect as the intriguing combination of coefficient bias and heteroscedasticity caused by the unknown scale factor. We argue that coefficient bias and heteroscedasticity associated with a chosen model can be effectively identified and assessed using two model selection criteria, namely, the *Average Absolute Values of Studentized Residuals* (A_k) and the *Relative Total Prediction Error* (R_k) for stratified firm groups. Our proposed deflator and model selection process consists of five major steps: (1) choose a working model which best reflects our understandings of the sample data and also meets our inference objectives of the study; (2) create a pool of candidate models based on the working model, including those which can be justified either statistically or economically; (3) stratify sampled firms into groups based on a chosen size measure of the firms; (4) evaluate each candidate model in the pool by computing the values of A_k and R_k for all size groups; and (5) select the best model from the pool by comparing A_k and R_k among candidate models.

Choosing an appropriate working model has never been an easy task for real world applications. It requires careful preliminary exploration of the data; it may also involve variable selection and/or other formal statistical procedures. The creation of the pool of candidate models is central to our proposed strategy. The pool needs to be large enough such that the best model identified from the pool is nearly free of the scale effect. The unknown scale factor can be viewed as a size variable which affects some or all other variables in a multiplicative way. We consider three types of candidate models for the pool including (i) deflated models with one of the observed independent variables as the deflator; (ii) deflated models with deflators depending on a single independent variable in a nonlinear fashion; and (iii) generalized linear models with model variances depending on the whole set of independent variables included in the working model. Without causing any confusion, we use GLM as abbreviation for 'Generalized Linear Modelling method' or 'Generalized Linear Model', depending on the circumstance.

Our study shows that the GLM methodology can be a valuable inference tool for market-based regression analyses. While deflation is potentially powerful in dealing with the 'scale effect' for cases where the underlying true scale factor is only related to a single independent variable, we argue that it is not suitable for cases where the scale factor depends on a set of independent variables. Deflation which

[1] Our proposed strategy includes the use of Generalized Linear Models (GLMs) and consequently involves model selections among the pool of candidate models.

may mitigate the scale effect for the former may cause spurious inference for the latter. Our simulation results reported in Sections III and IV indicate that the GLM can be an effective tool for mitigating the scale effect especially when the variance structure of the model is directly related to the mean value of the response variable. The use of GLM methodology as a tool for market-based regression analyses is briefly introduced in the next section. More details and computational notes are provided in the Appendix.

Our proposed approach does not contradict but rather complement prior studies on this topic. The 'best' deflated regression model can depend on several factors including the set of independent variables used in the model, the country or the region where the data are collected and the time period as well. Our view is consistent with Akbar and Stark (2003) in the sense that there is no universally 'best' deflator. What really important is that certain systematic treatments are followed and some unified criteria are used in identifying a model that exhibits the least 'scale effect'. Deflators used and models identified by prior studies can be naturally assessed and compared against each other using our proposed general statistical framework.

This study extends the existing literature in three main ways. First, we establish an objective, systematic, standardized and statistically meaningful way to identify an appropriate model which is least affected by the scale among alternative candidate models. Early studies (e.g. Christie, 1987) already documented the importance of this research problem and called for more appealing solutions. Our study answers the call and provides a general framework to address this important issue. Second, we argue and show that the average absolute values of studentized residuals and the relative total prediction error for firm size groups are valid criteria for identifying the existence of the scale effect and further measuring the severity of the effect. Third, we consider the complicated but practically important scenario where the scale factor is associated with more than one independent variable, and introduce the GLM methodology as an alternative 'scale effect' control mechanism under such situations.

We present the proposed general statistical framework for deflator and model selection in Section II, followed by an illustration of the proposed strategy in Section III through a simulation study using datasets which mimic the real data collected from COMPUSTAT. However, our simulation study is designed in such a way that the true model which generates the original scale-free variables and the true scale factor which leads to the 'actual' scale-affected observations are known under the simulation setting. Issues such as misspecified working models and different stratification schemes for the sampled firms are investigated in Section IV through a sensitivity analysis. Discussions and additional remarks are given in Section V. Some detailed description and computational notes on the GLM estimation theory are presented in the Appendix.

II. A Statistical Framework for Deflator and Model Selection

In this section, we first provide a brief discussion on 'scale' and 'scale-effect' and their impact on market-based regression analyses. Our main focus is to establish a general statistical framework for deflator and model selection. Major components of our proposed strategy include the choice of a working model, the creation of a pool of candidate models, criteria for the assessment of candidate models and the identification and selection of a model which is least affected by the scale effect.

Scale and scale effect

It is critical to our study to clarify 'scale' and 'scale-effect'. The two terms do not seem to have universally accepted definitions. Barth and Kallapur (1996) define 'scale' as an unobserved size variable S which has a multiplicative effect on all observed economical variables. More formally, let $Y_i^*, X_{1i}^*, \ldots, X_{pi}^*$, be the variables of interest associated with the unknown scale variable S_i for the i-th firm. What researchers actually observe are the scale affected variables $Y_i = S_i Y_i^*, X_{1i} = S_i X_{1i}^*, \ldots, X_{pi} = S_i X_{pi}^*$. Suppose that the relation between the true but unobserved variables is given by

$$Y_i^* = \beta_0^* + \beta_1^* X_{1i}^* + \cdots + \beta_p^* X_{pi}^* + e_i^* \qquad (1)$$

where the error terms e_i^* have zero mean and constant variance. The working regression model based on observed variables is specified as

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi} + e_i \qquad (2)$$

By multiplying S_i on both sides of (1) we obtain

$$Y_i = \beta_0^* S_i + \beta_1^* X_{1i} + \cdots + \beta_p^* X_{pi} + S_i e_i^* \qquad (3)$$

The two models (1) and (3) are mathematically equivalent but models (2) and (3) are not, since the unknown scale variable S_i is generally not available for inclusion in the working model (2).

There are two major implications for regression analysis in the presence of the scale variable: *Heteroscedasticity* and *Coefficient Bias*. Suppose that the true model (1) has a homogeneous variance structure for the error term, i.e. $\mathrm{Var}(e_i^*) = \sigma^2$, then the variance structure for the working model (2) becomes $\mathrm{Var}(e_i) = S_i^2\sigma^2$, which is nonhomogeneous due to the unequal size measure S_i. In addition, estimators of $\beta_j\,(j = 0, 1, \ldots, p)$ obtained from the working model (2) are typically biased for the true regression coefficients β_j^*. In the case of single independent variable X^*, the magnitude of the bias depends on the coefficient of variation related to X^* and S (Barth and Kallapur, 1996). In this article, we refer to the intriguing combination of heteroscedasticity and coefficient bias as the 'scale-effect'.

While heteroscedasticity is inherent to the scale and the scale effect, the concept of coefficient bias is quite different. Its existence and severity depend on whether the true model with all relevant variables is used for analysis or not. Statistical methods which are capable of handling heteroscedasticity do not necessarily correct for coefficient bias. For instance, in the case of known scale variable S_i and $\mathrm{Var}(e_i) = S_i^2\sigma^2$, heteroscedasticity can easily be handled by using the Weighted Least Square (WLS) method, with S_i^2 being the weights. This is equivalent to fitting the following model

$$\frac{Y_i}{S_i} = \beta_0\left(\frac{1}{S_i}\right) + \beta_1\left(\frac{X_{1i}}{S_i}\right) + \cdots + \beta_p\left(\frac{X_{pi}}{S_i}\right) + e_i \quad (4)$$

using the Ordinary Least Square (OLS) method. The coefficient bias problem may still persist since model (4) does not match (1) for the first term. White (1980) and MacKinnon and White (1985) propose several direct methods for dealing with heteroscedasticity even if the variance structure of the model is unknown. These methods, however, are not very useful to entirely mitigate the scale effect since they usually do not correct for coefficient bias.

Barth and Kallapur (1996) suggest two remedies for problems induced by the scale effect: deflate the original model by a selected scale proxy and/or include the scale proxy as an independent variable in the deflated model. It is apparent that including additional independent variables into the working model does not change the variance structure of the error terms, but it may help to remove coefficient bias. If the scale variable S_i is among the variables being observed in the sample data, then S_i itself should be free from scale affection. In other words, what we observe is still S_i, not $S_i \times S_i$. If S_i is included in the working model as an independent variable, this same variable should also be included in the deflated model, i.e., S_i should not be deflated by S_i, otherwise the variable S_i will be removed from the model. If the working model (2) includes all the independent variables involved in the true model (1) and if the scale variable S is known, then deflation simultaneously cures coefficient bias and heteroscedasticity and is the best remedy for problems induced by the scale effect.

The article by Easton and Sommers (2003) deserves some special comments and discussion. They seem to take a different view when considering the scale and the scale effect. They argue from an economic point of view that the 'true' scale is market capitalization, and define the scale effect as the overwhelming influence of large firms on regression estimation (Easton and Sommers, 2003, pp. 26, 41 and 51) due to large firms being large in both market capitalization and various accounting and economic variables. They identify the so-defined 'scale-effect' through studentized residuals obtained by fitting the following model (Easton and Sommers, 2003, p. 33, Equation 1)

$$\mathrm{MC}_j = \alpha_0 + \alpha_1\mathrm{BV}_j + \alpha_2\mathrm{NI}_j + \varepsilon_i \quad (5)$$

where MC_j is the market capitalization, BV_j is the book value of common equity and NI_j is the net income for the j-th firm. They use the WLS method to estimate the model parameters with weights being the square of market capitalization. This is equivalent to fitting the following model

$$1 = \alpha_0\frac{1}{\mathrm{MC}_j} + \alpha_1\frac{\mathrm{BV}_j}{\mathrm{MC}_j} + \alpha_2\frac{\mathrm{NI}_j}{\mathrm{MC_j}} + e_j \quad (6)$$

using the OLS method. From the computational point of view this approach is easy to implement, but the interpretation of results from fitting the model (6) is not straightforward under the traditional framework of regression analyses.

Easton and Sommers' (2003) approach, however, does raise an important research question: what is the appropriate strategy in regression analysis when the scale variable is closely related to the response variable? Note that $E(Y_i) = \beta_0 + \beta_1 X_i + \cdots + \beta_p X_{pi}$ under model (2), the question can be reiterated as: what is the appropriate approach when the scale variable is related to a linear combination of several independent variables? We argue in the section 'Candidate models' that the GLM method provides a possible solution to this scenario. Further investigation on the usefulness and effectiveness of the GLM methodology for market-based regression analyses is clearly needed.

Candidate models

We start with the working model (2) and term it as the baseline model. The choice of the response variable Y and the inclusion of independent variables for the baseline model depend on the economic background and inference objectives of a particular study. This is the crucial initial step for our proposed statistical framework but will not be addressed further in the current article. Our primary goal in this study is to search for an appropriate scale proxy such that the corresponding deflated model based on (2) is least affected by the scale assuming that the baseline model (2) is suitably chosen. To achieve this, we first create a pool of candidate models based on the chosen working model and then identify and select the best model from the pool using appropriate criteria. The pool should include models which are economically and/or statistically plausible and meaningful. Economic considerations often justify the inclusion or exclusion of certain variables as potential deflator candidates. Without having such an economic background, we consider models which are compatible with models (1) and (2).

The assumed models (1) and (2) imply that $\mathrm{Var}(Y_i|X_i, S_i) = V_i\sigma^2$ with $V_i = S_i^2$. This provides an effective way of choosing deflator candidates based on the variance structure. There are two commonly encountered variance structures of the error term in practice:

(i) V_i can be approximated by $|X_{ji}|^r$ for some j and $r > 0$, where X_{ji} is the j-th independent variable associated with the i-th firm. In other words, V_i depends only on a single independent variable.
(ii) V_i is related to a certain combination of several independent variables.

For scenario (i), the most sensible choice of r is either 1 or 2, i.e. $V_i = |X_{ji}|$ or $V_i = X_{ji}^2$. This leads to deflating the baseline regression model by $\sqrt{|X_j|}$ when r is 1 or by X_j when r is 2. Most prior market-based studies consider scenario (i) and assume $r = 2$ (see, for instance, Barth and Kallapur, 1996; Lo and Lys, 2000; Akbar and Stark, 2003; Easton and Sommers, 2003, among others). In this article we consider both $r = 1$ and $r = 2$ but other choices of r are also possible. It is important to note that the scale variable S_i, although unobserved, can be viewed as a size measure of the firms. Observed independent variables which might be interpreted as a size measure, such as the book value of common equity, the net income, sales, the number of shares outstanding or even the market capitalization, should be considered as potential deflator candidates and the corresponding deflated models should be included in the pool.

Situations under scenario (ii) are somewhat complicated. The conventional deflation method using a single independent variable as deflator is not appropriate under this scenario and can sometimes cause even more severe coefficient bias, resulting in invalid and spurious inference. This situation should be dealt with the GLM method as discussed below.

Let $\mu_i = E(Y_i|\mathbf{X}_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$ be the mean function under the linear regression model (2.2) and $\mathrm{Var}(Y_i|\mathbf{X}_i) = V_i\sigma^2$ be the variance structure, where $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})$. In theory, V_i may depend on the whole set of independent variables $\mathbf{X}_i = (X_{1i}, \ldots, X_{pi})$. A good approximation for many statistical applications is that the V_i relates to $\mathbf{X}_i$ only through the mean function μ_i. This is true for many business applications where firms with larger mean values will also exhibit larger variations. Scatter plots of the COMPUSTAT data used for our simulation studies display a clear trend of this feature. One possible way to specify this type of relationship is to use a power function $V_i = \mu_i^r$ for some $r > 0$.

One of the major statistical advances in the past 15 years is the development of the GLM theory. The method is more general than the traditional OLS or WLS approach, and is extremely powerful in dealing with cases where V_i depends on $\mathbf{X}_i$ only through μ_i. We therefore propose to use the GLM method under scenario (ii) assuming $V_i = \mu_i^r$. It should be noted that GLM is not deflation. It is most effective in handling heteroscedasticity under the current situation but its effectiveness in dealing with the coefficient bias problem requires further investigation. In Sections III and IV, we show through simulation studies that the GLM method can be effective for both, and is potentially a very useful inference tool for market-based regression analyses. Some technical as well as computational details about the GLM method under the current context are presented in the appendix.

In our simulation studies, the pool of candidate models includes (i) deflated regression models with deflators chosen from the set of independent variables; (ii) deflated regression models using square roots of the variables used in (i) as deflators; and (iii) two GLMs with $V_i = |\mu_i|$ or $V_i = \mu_i^2$ where $\mu_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_p X_{pi}$. Those two variance structures are also popular choices for the GLM method in other areas of applications. We assess

each model using our proposed criteria, i.e. *Goodness-of-Fit* and *Prediction Power* discussed in the next section, for different size groups.

Goodness-of-fit and prediction power

The best model to be selected from the pool of candidate models should be the one which is least affected by the scale. Ideally, one should assess these models by conducting formal tests to see whether heteroscedasticity and/or coefficient bias are removed or mitigated for a particular candidate model. Such formal tests, however, are either undesirable or unavailable under current situations. First, there exist a large number of plausible candidate models, including those obtained through deflation, which can be justified by either economical arguments or statistical considerations. It is very likely that more than one candidate model can survive for a test on the homogeneity of the error terms, but the problem of selecting the 'best' model remains unsolved. Second, the issue of coefficient bias is very subtle. The true coefficient bias is directly related to whether the candidate model is identical or very similar to the true model, which is not testable for any practical situations. The two major aspects of the scale effect, namely heteroscedasticity and coefficient bias, however, can be examined in an indirect way through the goodness-of-fit of the data to the chosen model and the prediction power of the model for different firm size groups. We elaborate the three involved components below.

(1) Why stratify firms by a size variable?

Sample data collected from balance sheets and income statements differ significantly by order of magnitude, depending on the size of the firm. On the other hand, the so-called scale variable, neither explicitly defined nor precisely observed, must be positively correlated to other firm size variables. If we group firms by a size variable and evaluate the performance of each candidate model across all firm size groups, the scale effect, if exists, will show up through the 'nonuniform' model behaviour across groups of different sizes. In our proposed deflator and model selection strategies, we divide sample firms into groups based on a chosen size measure, and assess the model behaviour and performance within each of these groups and compare them across the groups. A model which is free from the scale effect will show uniform performance in terms of the average absolute values of studentized residuals and the relative total prediction error (to be defined in the sequel) across all size groups.

There are two other issues related to grouping: (i) which size variable should we use if there are multiple possible choices? and (ii) how many groups are appropriate to effectively identify the scale effect? Those issues are addressed and some empirical evidences are obtained in our simulation and sensitivity studies reported in Sections III and IV.

(2) How to check heteroscedasticity?

There exist several methods which can be used to directly deal with heteroscedasticity (see, for instance, White (1980) and MacKinnon and White (1985)). However, we are not interested in those procedures since our goal is to simultaneously handle both heteroscedasticity and coefficient bias. Under the assumption that the pool of candidate models includes one which is free from the scale effect, the real question of interest here is how to identify a model with a homogeneous variance structure and free of coefficient bias as well. Studentized residuals obtained from fitting a chosen model can be used to identify the existence and to further measure the severity of heteroscedasticity of the model.

Suppose sample firms are stratified into K groups, based on a chosen firm size measure. Let G_k be the set of firms and m_k be the number of firms in the k-th group, $k = 1, 2, \ldots, K$; let $\hat{\varepsilon}_i$ be the studentized residuals for the i-th firm obtained from fitting the chosen candidate model using OLS. The *Average Absolute Values of the Studentized Residuals* for the k-th group is defined as

$$A_k = \frac{1}{m_k} \sum_{i \in G_k} |\hat{\varepsilon}_i|$$

This quantity summarizes the *goodness-of-fit* of the model to the sample data for the k-th group. If the candidate model has a homogeneous variance structure and if the total sample size is large, then the $\hat{\varepsilon}_i$'s follow approximately a standard normal distribution and are also approximately independent of each other. Consequently,

$$E(|\hat{\varepsilon}_i|) \doteq \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \mathrm{d}x = \sqrt{\frac{2}{\pi}} \doteq 0.8$$

and $\mathrm{Var}(|\hat{\varepsilon}_i|) \doteq E(\hat{\varepsilon}_i^2) - 0.8^2 \doteq 1 - 0.64 = 0.36$. A one-sided 95% level upper prediction bound for A_k is given by $0.8 + 1.645\sqrt{0.36/m_k}$. For instance, if $m_k = 100$, then this upper bound is 0.899. Cases where one or several values of A_k go beyond the upper bound should be viewed as evidence of heteroscedasticity associated with the candidate model. The severity of heteroscedasticity is therefore measured by the magnitude of A_k exceeding the

upper bound. On the other hand, a model with uniformly smaller values of A_k should be considered as more desirable when we try to select the best model from the pool.

(3) How to assess coefficient bias?

The assessment for coefficient bias is the most difficult and subtle aspect for deflator and model selection. As we argued in previous sections, coefficient bias is neither identifiable nor estimable. It is essentially the question of whether the candidate model is the same as the underlying true model, which can never be answered for sure based on a particular sample dataset.

There are two fundamental objectives for regression analysis: to describe how the response variable (or dependent variable) is related to important independent variables (or covariates) using observed data, and to make predictions for 'future' values using the estimated regression equation. While a larger-than-necessary model which includes redundant independent variables can sometimes provide improved goodness-of-fit to a particular dataset, it is usually less powerful for predicting new cases and future values. Under the current context, if a candidate model is free of the scale effect, i.e. it is close to the underlying true model, than it should demonstrate a good balance between goodness-of-fit and prediction power. To assess the existence of coefficient bias in a candidate model, we therefore require a measure of prediction power of the model. A model with poor prediction capacity implies that the model is not close to the underlying true one, which further implies that coefficient bias may exist. The prediction power together with the goodness-of-fit criterion as measured by the *Average Absolute Values of the Studentized Residuals* will provide indirect but valid assessment for coefficient bias.

Let D_i be the chosen candidate deflator. The deflated model is given by

$$\frac{Y_i}{D_i} = \beta_0 + \beta_1\left(\frac{X_{1i}}{D_i}\right) + \cdots + \beta_p\left(\frac{X_{pi}}{D_i}\right) + e_i \qquad (7)$$

To assess the prediction power of model (7), we propose to use the following procedure which is similar to the cross-validation technique often used in statistics. Let $(Y_i, X_{1i}, \ldots, X_{pi}, D_i)$ be the i-th row of the data matrix associated with the i-th firm in the sample. We fit model (7) using the OLS method and the dataset with the i-th row removed from the sample. Let $(\hat{\beta}_0[-i], \hat{\beta}_1[-i], \ldots, \hat{\beta}_p[-i])$ be the estimated regression coefficients, where $[-i]$ indicates that the i-th firm is deleted from the sample. The predicted value of Y_i is then computed as

$$\hat{Y}_i = \hat{\beta}_0[-i]D_i + \hat{\beta}_1[-i]X_{1i} + \cdots + \hat{\beta}_p[-i]X_{pi}$$

This procedure is carried out for every firm included in the sample. The prediction error for the i-th firm is given by $Y_i - \hat{Y}_i$. It should be noted that this error does not belong to any type of residuals, since Y_i is not used in finding $\hat{Y}_i$.

The total prediction error $\sum_{i\in G_k} |Y_i - \hat{Y}_i|$ for each of the K groups may be used to measure the prediction power of the candidate model. This measurement, however, is difficult to interpret when it comes to comparisons between firm groups of different sizes. For instance, the error $|Y_i - \hat{Y}_i| = 0.1$ should be viewed as unacceptably large if the true value is $Y_i = 0.001$; on the other hand, the error $|Y_i - \hat{Y}_i| = 1000$ is negligible if $Y_i = 1\,000\,000$. We therefore propose to use the *Relative Total Prediction Error* for each of the K size groups, denoted by R_k and defined below for the k-th group as the measure of prediction power of the model,

$$R_k = \frac{\sum_{i\in G_k} |Y_i - \hat{Y}_i|}{\sum_{i\in G_k} |Y_i|}$$

If the model is acceptable, the R_k's must be comparable to each other among different firm size groups. In addition, smaller values of R_k indicate stronger prediction power of the model.

For GLMs, the estimated regression coefficients $(\hat{\beta}_0[-i], \hat{\beta}_1[-i], \ldots, \hat{\beta}_p[-i])$ are obtained through the quasi-score method (see the Appendix for details). The predicted value of Y_i is simply computed as $\hat{Y}_i = \hat{\beta}_0[-i] + \hat{\beta}_1[-i]X_{1i} + \cdots + \hat{\beta}_p[-i]X_{pi}$. As indicated previously, no deflator is involved in the GLM approach.

III. Simulation Study

We illustrate the proposed deflator and model selection process and demonstrate the effectiveness of our proposed approach through a simulation study. The data we use for the study mimic real world applications but both the true regression model and the true scale variable are set to be known under the simulation setting. Our simulation study is programmed using the free statistical software package *R* which is downloadable at www.r-project.org (R Development Core Team, 2005). The original datasets and the *R* programs used in the simulation are available from the first author of this article upon request.

The dataset based on which our simulation samples are generated is collected from COMPUSTAT for the year 1999. The data file contains complete information on market value of equity (MV), book value of equity (BV), net income (NI), R&D spendings (RD), sales (SA), cash flows (NFO), total assets (TA), dividends (DI), capital contributions (CC) and number of shares (SH) outstanding. The total number of firms with these types of information available is 3713. Since this is a real dataset, we treat it as if all its observations are affected by an unknown scale variable. To create a dataset which is 'scale-free', we sort the data file by SA,[2] and the first $N = 3240$ observations from the sorted file are used to form the 'scale-free' dataset. The motivation behind throwing away a portion of large observations is that we will re-create these large values through a chosen scale variable.

To ensure that sample data used for our simulation studies follow a known regression model, we generate the response variable Y_i^* using the following model,

$$Y_i^* = \beta_0^* + \beta_1^* \mathrm{BV}_i^* + \beta_2^* \mathrm{NI}_i^* + \beta_3^* \mathrm{RD}_i^* + e_i^* \qquad (8)$$

where BV*, NI* and RD* are the 'scale-free' variables and the error terms e_i^* follow an $N(0, \sigma^2)$ distribution. The regression coefficients $\beta_j^* = (j = 0, 1, 2, 3)$ are chosen as the estimated regression coefficients by fitting the model (8) using MV* as the response variable. In doing so the generated response variable Y^* mimics MV* but it follows the known model (8); the error variance σ^2 is chosen such that the sample multiple correlation coefficient between Y_i^* and the set of independent variables (BV_i^*, NI_i^*, RD_i^*) is 0.80; an constant number is added to all Y_i^* so that min $Y_i^* = 0.02$ (this is the minimum value of MV* from the original dataset). Our final 'scale-free' dataset consists of $N = 3240$ observations on variables Y^*, BV*, NI*, RD*, DI*, CC*, SA* and MV*.

The scale-affected 'observed' sample data are generated through the true scale variable. Under model (8), the mean response values are given by $\mu_i = \beta_0^* + \beta_1^* \mathrm{BV}_i^* + \beta_2^* \mathrm{NI}_i^* + \beta_3^* \mathrm{RD}_i^*$. They are known under the simulation setting. We consider four scenarios for the true scale variable S. Those are also likely representative cases in practice.

(i) $S_i = \mathrm{BV}_i$: the scale is one of the observed independent variables.
(ii) $S_i = \sqrt{\mathrm{BV}_i}$: the scale is related to but not the same as one of the observed independent variables.
(iii) $S_i = \mu_i$: the scale is linearly related to the set of independent variables in the true model.
(iv) $S_i = \sqrt{\mu_i}$: the scale nonlinearly relates to the set of independent variables.

Under each scenario, the scale-affected observations are obtained by multiplying the scale-free variables by the scale variable. In cases (1) and (2), the scale variable BV itself remains unaffected. Four datasets are generated, one for each of the four scenarios, and are used as 'observed' sample data in the simulation study.

In what follows, we go through the process according to our proposed strategy to select a deflator and/or identify a model which is least affected by the scale. The first step is to choose a working model. We consider the following baseline model:

$$Y_i = \beta_0 + \beta_1 \mathrm{BV}_i + \beta_2 \mathrm{NI}_i + \beta_3 \mathrm{RD}_i + e_i \qquad (9)$$

This model has the same structure as the true one. The practically important issue of misspecified working models will be explored in the next section.

The second step is to create a pool of candidate models. We consider three types of models: (i) deflation models using the two independent variables BV and NI and their square roots as deflators; (ii) deflated models with the exogenous variable MV and its square root as deflators; and (iii) the two GLMs with the variance function (a) $V_i = \mu_i$ and (b) $V_i = \mu_i^2$, denoted as GLM_a and GLM_b. The variable RD is weakly correlated to the response variable and is not considered as a sensible choice of the scale variable. Our final pool of candidate models can be represented by the set

$$\Omega = \left\{ \mathrm{U}, \mathrm{BV}, \mathrm{NI}, \mathrm{MV}, \sqrt{\mathrm{BV}}, \sqrt{\mathrm{NI}}, \sqrt{\mathrm{MV}}, \mathrm{GLM}_a, \mathrm{GLM}_b \right\}$$

where 'U' represents the undeflated working model (9). There are a total number of nine candidate models included in the pool.

The third step is to group the sample firms by a size measure. There are several variables which can be interpreted as possible size measures. We consider sales (SA) in this section but other choices are also examined in Section IV. The number of groups is another important issue. We report results for $K = 40$ groups in this section and discuss more on alternative grouping schemes in Section IV. The number of firms in each group is $3240/40 = 81$ when $K = 40$.

[2] This is our size measure for simulation results reported in this article. Other choices of size measure are also considered, and our major conclusions remain unchanged under different choices of size variables.

Table 1. Average absolute values of studentized residuals for groups 1–4 and 37–40

Scale	Model	G1	G2	G3	G4	G37	G38	G39	G40
BV_i	U	0.39	0.38	0.38	0.37	0.92	1.04	1.24	2.19
	BV	0.79	0.83	0.83	0.81	0.72	0.74	0.84	0.77
	NI	0.64	1.25	1.01	0.90	0.11	0.11	0.09	0.09
	MV	2.10	2.12	1.42	1.34	0.05	0.06	0.06	0.06
	$\sqrt{BV}$	0.45	0.41	0.36	0.35	0.79	0.91	1.01	2.07
	$\sqrt{NI}$	0.25	0.43	0.50	0.50	0.53	0.67	0.68	1.43
	$\sqrt{MV}$	0.49	0.74	0.78	0.79	0.67	0.77	0.78	1.21
	GLM_a	0.25	0.14	0.13	0.10	0.69	0.87	0.91	2.01
	GLM_b	0.81	0.89	0.89	0.85	0.56	0.63	0.61	0.87
$\sqrt{BV_i}$	U	0.67	0.52	0.46	0.40	1.20	1.12	1.60	2.09
	BV	2.32	0.76	0.51	0.54	0.41	0.37	0.35	0.49
	NI	2.11	1.46	0.71	0.92	0.16	0.17	0.14	0.14
	MV	2.66	0.79	0.52	0.41	0.07	0.07	0.06	0.05
	$\sqrt{BV}$	0.73	0.86	0.79	0.81	0.76	0.84	0.84	0.96
	$\sqrt{NI}$	1.25	1.39	0.93	1.23	0.46	0.49	0.54	0.88
	$\sqrt{MV}$	1.73	1.55	1.25	1.25	0.73	0.78	0.80	0.79
	GLM_a	1.24	0.88	0.76	0.64	0.96	0.86	1.17	1.10
	GLM_b	1.62	0.98	0.83	0.78	0.68	0.69	0.74	0.67
μ_i	U	0.61	0.62	0.68	0.65	1.11	1.09	1.26	1.68
	BV	0.99	0.09	0.02	0.17	0.02	0.02	0.03	0.02
	NI	1.21	0.98	0.29	0.71	0.14	0.10	0.08	0.10
	MV	1.11	0.81	0.67	0.83	0.42	0.26	0.27	0.23
	$\sqrt{BV}$	1.96	0.80	0.51	0.54	0.18	0.15	0.18	0.41
	$\sqrt{NI}$	1.80	1.45	0.85	1.00	0.32	0.30	0.28	0.75
	$\sqrt{MV}$	1.43	1.08	0.92	1.05	0.71	0.69	0.81	1.20
	GLM_a	0.73	0.75	0.85	0.81	0.85	0.82	0.91	0.99
	GLM_b	0.83	0.84	0.98	0.92	0.59	0.61	0.61	0.54
$\sqrt{\mu_i}$	U	0.69	0.70	0.80	0.78	0.85	0.81	1.18	1.05
	BV	0.94	0.07	0.02	0.18	0.02	0.02	0.02	0.02
	NI	1.10	0.90	0.25	0.73	0.10	0.10	0.08	0.10
	MV	1.12	0.82	0.66	0.83	0.30	0.38	0.26	0.25
	$\sqrt{BV}$	1.86	0.72	0.44	0.49	0.18	0.16	0.19	0.37
	$\sqrt{NI}$	1.76	1.42	0.78	1.00	0.33	0.34	0.31	0.68
	$\sqrt{MV}$	1.50	1.14	0.94	1.12	0.74	0.74	0.83	1.19
	GLM_a	0.77	0.78	0.90	0.88	0.72	0.64	0.90	0.70
	GLM_b	0.84	0.85	0.98	0.95	0.60	0.49	0.68	0.46

The fourth step is to obtain numeric results. For each of the four datasets, the average absolute values of studentized residuals A_k and the relative total prediction errors R_k are computed for all 40 groups under each of the nine candidate models in the pool. The computation of R_k is quite intensive since we need to fit the candidate model using the delete-1 dataset for each firm, and the process needs to be repeated for all firms in the sample as well as for all candidate models.

A brief review of values of A_k and R_k for all 40 groups reveals an interesting phenomenon which may have practically important implications: if the values of A_k or R_k display a nonuniform pattern over the 40 groups, the larger or smaller values are always found in the first or last four groups. We therefore only report results for the first four groups (G1–G4) and the last four groups (G37–G40) for A_k (Table 1) and R_k (Table 2). Values of A_k and R_k for the unreported groups are always somewhere in-between and hence are not very informative for our decision makings.

The last step is to identify the 'best' model from the pool of candidate models using the criteria of goodness-of-fit and prediction power measured by A_k and R_k. To make comparisons among different candidate models considered in the simulation, we note that (i) the 95% level upper prediction bound for A_k is $0.8 + 1.645 \times 0.6/\sqrt{81} \doteq 0.91$ for $m_k = 81$ (see the section 'Goodness-of-fit and prediction power' for detailed argument), and large values of A_k are viewed as evidence of heteroscedasticity; and (ii) a nonuniform pattern displayed by the relative total prediction error R_k over different

Table 2. Relative total prediction errors for groups 1–4 and 37–40

Scale	Model	G1	G2	G3	G4	G37	G38	G39	G40
BV_i	U	97.3	34.1	24.1	17.0	0.34	0.30	0.34	0.25
	BV	0.21	0.27	0.16	0.24	0.09	0.12	0.09	0.07
	NI	0.22	0.29	0.16	0.24	0.61	0.54	0.70	1.50
	MV	0.20	0.27	0.16	0.24	1.37	0.49	0.65	1.15
	$\sqrt{BV}$	2.15	1.22	0.99	0.79	0.11	0.17	0.12	0.21
	$\sqrt{NI}$	0.28	0.28	0.17	0.25	0.11	0.18	0.15	0.25
	$\sqrt{MV}$	0.21	0.27	0.16	0.24	0.11	0.19	0.12	0.23
	GLM_a	0.50	0.29	0.18	0.23	0.10	0.18	0.12	0.22
	GLM_b	0.20	0.27	0.16	0.24	0.11	0.19	0.12	0.24
$\sqrt{BV_i}$	U	2.29	1.25	0.78	0.79	0.13	0.13	0.10	0.09
	BV	0.70	0.75	0.61	0.67	3.46	4.36	4.04	10.2
	NI	0.60	0.39	0.43	0.42	7.07	10.8	7.19	20.2
	MV	0.60	0.80	0.98	1.00	7.68	9.27	10.7	13.1
	$\sqrt{BV}$	0.24	0.21	0.24	0.22	0.12	0.12	0.12	0.13
	$\sqrt{NI}$	0.45	0.34	0.36	0.37	0.46	0.46	0.48	0.79
	$\sqrt{MV}$	0.61	0.48	0.53	0.47	0.66	0.72	0.74	0.71
	GLM_a	1.68	0.84	0.52	0.50	0.14	0.14	0.12	0.11
	GLM_b	0.97	0.43	0.30	0.27	0.24	0.25	0.30	0.30
μ_i	U	0.24	0.20	0.20	0.24	0.17	0.12	0.11	0.09
	BV	1.20	1.06	1.43	2.03	48.1	45.7	53.1	268
	NI	0.83	0.64	0.70	0.85	26.0	25.8	22.9	95.4
	MV	5.32	4.68	5.65	6.95	185	158	171	279
	$\sqrt{BV}$	0.53	0.42	0.38	0.41	1.07	0.67	1.15	2.24
	$\sqrt{NI}$	0.51	0.44	0.41	0.44	1.38	1.24	1.18	2.61
	$\sqrt{MV}$	0.75	0.73	0.71	0.66	3.62	3.42	4.20	4.90
	GLM_a	0.23	0.20	0.19	0.24	0.15	0.10	0.11	0.12
	GLM_b	0.23	0.20	0.19	0.24	0.14	0.10	0.13	0.18
$\sqrt{\mu_i}$	U	0.23	0.20	0.20	0.23	0.14	0.12	0.12	0.07
	BV	1.36	1.29	1.73	2.47	51.4	46.6	54.1	29.5
	NI	0.83	0.63	0.71	0.91	27.1	33.2	22.5	85.5
	MV	5.37	4.75	6.03	7.30	148	186	138	256
	$\sqrt{BV}$	0.51	0.40	0.37	0.39	1.32	1.04	1.43	2.95
	$\sqrt{NI}$	0.51	0.43	0.42	0.43	1.51	1.76	1.38	3.16
	$\sqrt{MV}$	0.79	0.78	0.75	0.68	4.29	4.42	4.79	6.37
	GLM_a	0.23	0.20	0.20	0.23	0.14	0.12	0.12	0.09
	GLM_b	0.23	0.20	0.20	0.23	0.14	0.12	0.12	0.10

firm size groups indicates that the model is poor or wrong, which further implies that the coefficient bias problem is of concern. The tolerable upper bound of R_k depends on how strongly the response variable is correlated to the set of independent variables. For most applications the prediction power should be viewed as strong if $R_k < 0.20$ for all size groups but this argument is quite arbitrary. The bottom line is that the smaller the R_k the stronger the prediction power.

Tables 1 and 2 contain results for the four datasets, each with a different true scale variable. Our major findings can be summarized as follows:

(i) For the scenario where $S_i = BV_i$: the deflated model with $D_i = BV_i$ and the GLM_b are the only acceptable models under the criteria A_k; the two models also have identical performance in terms of R_k for the first four groups but the deflated model is slightly better when judged by R_k for the last four groups.

(ii) For the case where $S_i = \sqrt{BV}_i$: the deflated model with $D_i = \sqrt{BV}_i$ is the only model with all A_k bounded by 0.96; this is also the best model under the criterion R_k.

(iii) For the third scenario where $S_i = \mu_i$: the two GLMs have satisfactory performance under both criteria; the deflated model with $D_i = BV_i$ seems working well in terms of goodness-of-fit but has outrageous performance for prediction; the baseline model itself (U) has good prediction power but fits the data poorly.

(iv) For the last case where $S_i = \sqrt{\mu_i}$: this case is very much similar to scenario (iii). The two

GLMs are indeed better than all others and both models are satisfactory.

It is also observed that the exogenous variable MV, based on which our response variable tries to mimic, is not a statistically meaningful deflator. In cases (iii) and (iv) where the true scale variable is related to the whole set of independent variables, none of the traditionally deflated models, i.e. using a single variable as deflator, is effective. Indeed all these deflated models produce spurious and in some cases outrageous results, as judged by the prediction capacity of the model. Note that $R_k = (\sum_{i\in G_k} |Y_i - \hat{Y}_i|)/(\sum_{i\in G_k} |Y_i|) = 1$ if we simply use $\hat{Y}_i = 0$ as predicted values for all sample firms, the model must be outrageously wrong if $R_k > 1$ for any firm size group. The GLMs, GLM_a and GLM_b, provide more efficient and reliable inference in cases (iii) and (iv), and the two models have similar and desirable behaviour under both assessing criteria.

IV. Sensitivity Analysis

For any real world application of the proposed deflator and model selection strategy, different decisions and choices could be made at each of the major steps as outlined in Section III. An incomplete list of these decisions includes (i) the choice of a size variable used for stratification; (ii) the number of groups (or strata) used for evaluation; (iii) the choice of a baseline working model; and (iv) candidate models to be included in the pool. In this section, we consider some of these aspects and conduct a sensitivity analysis. We focus here on (ii) and (iii) and see how things unfold if different grouping methods and/or misspecified working models are used. We will briefly summarize our findings regarding (i) and (iv) at the end of this section.

We first look at the issue of stratification, using the dataset with $S_i = \sqrt{\text{BV}_i}$ as an example. We compute the values of A_k and R_k under three grouping schemes with the number of groups being 10, 20 and 40, respectively. The variable SA is used as the size measure. Similar to what we observed in Section III, for each of the three grouping schemes, if values of A_k and R_k display a nonuniform pattern, the larger and smaller values are always shown in the first and last three or four groups. Table 3 presents the values of R_k for the first four groups (G1–G4) and the last four groups (G-4–G-1) under each of the three grouping schemes. There are two clear messages conveyed from the table: for the best model ($D_i = \sqrt{\text{BV}_i}$) the values of R_k remain virtually unchanged under different grouping schemes, but for models with R_k displaying a nonuniform pattern over the groups the nonuniformity is more pronounced under the more refined grouping scheme (i.e. 40 groups), which suggests that using a larger number of size groups can help better identify coefficient bias associated with the candidate model. This finding is also confirmed by other datasets considered in Section III and by the results of A_k (not reported here) as well.

Another major concern for all real world applications is the possible misspecification of a working model. While outrageously wrong models could be avoided by a careful selection of the response variable and the set of independent variables based on economic and statistical considerations, some mild departure of the working model from the true one can never be eliminated. The most common problems are either an over-specified working model with too many independent variables or an under-specified model with important covariates missing from the model.

We repeat the simulation study of Section III using the same datasets generated from model (8) but two different working models. The sample correlation coefficients between the response variable Y and the five independent variables in the order of BV, NI, RD, CC and DI are, respectively, 0.94, 0.85, 0.50, 0.44 and 0.22 for the dataset with $S_i = \sqrt{\text{BV}_i}$. This leads to the following natural consideration of under-specified and over-specified working models:

$$Y_i = \beta_0 + \beta_1 \text{BV}_i + e_i \tag{10}$$

$$Y_i = \beta_0 + \beta_1 \text{BV}_i + \beta_2 \text{NI}_i + \beta_3 \text{RD}_i + \beta_4 \text{CC}_i + \beta_5 \text{DI}_i + e_i \tag{11}$$

The correctly specified working model is (9) used in Section III. We compute the values of A_k and R_k under models (10) and (11) using 40 groups based on the firm size measure SA. Values of R_k for the first and last four groups are presented in Table 4, where the first column C–M denotes *Candidate Models* considered in the pool and the second column W–M represents *Working Models* (i.e. models (10) and (11)) used to produce the candidate model. Results under the correct working model (9) are repeated here for the purpose of comparison.

All candidate models produce comparable values of R_k for groups G1–G4 under all three working models but some behave quite differently for groups G37–G40. The most important observation

Table 3. Relative total prediction errors under different groupings

Model	#Group	G1	G2	G3	G4	G-4	G-3	G-2	G-1
U	10	1.13	0.44	0.27	0.22	0.19	0.17	0.15	0.11
	20	1.67	0.78	0.52	0.38	0.15	0.14	0.13	0.10
	40	2.29	1.25	0.78	0.79	0.13	0.13	0.10	0.09
BV	10	0.68	0.54	0.42	0.53	2.16	2.35	3.64	6.58
	20	0.73	0.64	0.56	0.52	3.89	3.47	3.92	8.13
	40	0.70	0.75	0.61	0.67	3.46	4.36	4.04	10.2
NI	10	0.44	0.52	0.77	1.50	5.19	5.53	6.08	13.2
	20	0.48	0.42	0.45	0.58	6.27	5.95	9.01	15.7
	40	0.60	0.39	0.43	0.42	7.07	10.8	7.19	20.2
MV	10	0.88	1.82	2.54	3.93	9.73	7.53	14.38	10.9
	20	0.72	0.99	1.25	2.30	12.3	15.8	8.49	12.3
	40	0.60	0.80	0.98	1.00	7.68	9.27	10.7	13.1
$\sqrt{\text{BV}}$	10	0.22	0.21	0.21	0.20	0.17	0.16	0.14	0.12
	20	0.22	0.23	0.22	0.21	0.16	0.13	0.12	0.12
	40	0.24	0.21	0.24	0.22	0.12	0.12	0.12	0.13
$\sqrt{\text{NI}}$	10	0.37	0.32	0.27	0.31	0.46	0.37	0.42	0.60
	20	0.38	0.36	0.35	0.29	0.46	0.39	0.46	0.68
	40	0.45	0.34	0.36	0.37	0.46	0.46	0.48	0.79
$\sqrt{\text{MV}}$	10	0.51	0.38	0.30	0.33	0.61	0.64	0.82	0.71
	20	0.53	0.50	0.40	0.37	0.85	0.80	0.69	0.72
	40	0.61	0.48	0.53	0.47	0.66	0.72	0.74	0.71
GLM_a	10	0.77	0.31	0.24	0.22	0.20	0.17	0.16	0.12
	20	1.18	0.51	0.36	0.27	0.17	0.15	0.14	0.12
	40	1.68	0.84	0.52	0.50	0.14	0.14	0.12	0.11
GLM_b	10	0.42	0.25	0.25	0.23	0.20	0.21	0.28	0.28
	20	0.64	0.28	0.26	0.23	0.27	0.28	0.24	0.30
	40	0.97	0.43	0.30	0.27	0.24	0.25	0.30	0.30

is that the relative performance of these candidate models remains the same under all three working models. For instance, the deflated model using $\text{D}_i = \sqrt{\text{BV}_i}$ is still the best while the two deflated models with $D_i = \text{NI}_i$ and $D_i = \text{MV}_i$ are among the worst.

We also explored the sensitivity of our proposed strategy using (i) COMPUSTAT data from 1997 and 1998 to generate sample data for the simulation; and (ii) number of shares (SH) outstanding and market value of equity (MV) as firm size measures. Under all the scenarios considered in the analysis, there are two common observations speaking loud and clear. First, if the true scale variable depends on a single independent variable and if the corresponding deflated model is included in the pool of candidate models, this deflator and the associated deflated model are always correctly identified. Second, if the true scale variable is related to the whole set of independent variables through the mean function, the two GLMs are always the best under the criteria A_k and R_k, but there seems to be no clear-cut winner between GLM_a and GLM_b.

V. Concluding Remarks

Deflator and model selection for market-based regression analyses involves both economical and statistical considerations. For most level-based research designs, one of the important econometrical tasks is to control and mitigate the 'scale effect'. Most prior studies on this topic are conducted by accounting researchers. Those studies focus on economic justifications for individual deflator candidate using various baseline models. Conclusions and recommendations are often restricted to particular datasets and specific baseline models used in the study. No generalizable deflator in an economic sense has been identified.

The current study establishes a unified statistical framework to guide the deflator and model selection process. Given the fact that the true 'scale' and the magnitude of 'scale effect' are both unknown, the proposed framework provides an objective, systematic and statistically meaningful way to select an appropriate model which is least affected by the 'scale'. The proposed framework consists of two integrated parts: The creation of a

Table 4. Relative total prediction errors under different working models

C–M	W–M	G1	G2	G3	G4	G37	G38	G39	G40
U	(10)	2.29	1.26	0.79	0.80	0.24	0.21	0.18	0.19
	(11)	2.29	1.25	0.78	0.79	0.13	0.13	0.11	0.09
	(9)	2.29	1.25	0.78	0.79	0.13	0.13	0.10	0.09
BV	(10)	0.54	0.40	0.44	0.38	1.01	1.28	1.11	4.11
	(11)	0.68	0.73	0.59	0.66	3.53	4.51	4.38	10.0
	(9)	0.70	0.75	0.61	0.67	3.46	4.36	4.04	10.2
NI	(10)	0.59	0.40	0.43	0.41	10.5	11.8	11.4	14.4
	(11)	0.59	0.38	0.44	0.41	6.30	9.46	6.57	17.9
	(9)	0.60	0.39	0.43	0.42	7.07	10.8	7.19	20.2
MV	(10)	0.60	0.64	0.88	0.98	25.0	30.8	31.6	42.2
	(11)	0.55	0.74	0.98	0.96	8.33	7.86	11.0	13.3
	(9)	0.60	0.80	0.98	1.00	7.68	9.27	10.7	13.1
$\sqrt{\text{BV}}$	(10)	0.24	0.22	0.25	0.22	0.24	0.21	0.18	0.21
	(11)	0.24	0.21	0.24	0.22	0.12	0.12	0.11	0.13
	(9)	0.24	0.21	0.24	0.22	0.12	0.12	0.12	0.13
$\sqrt{\text{NI}}$	(10)	0.52	0.46	0.43	0.42	0.59	0.71	0.71	0.63
	(11)	0.45	0.34	0.36	0.37	0.41	0.44	0.48	0.77
	(9)	0.45	0.34	0.36	0.37	0.46	0.46	0.48	0.79
$\sqrt{\text{MV}}$	(10)	0.69	0.60	0.59	0.55	0.83	1.01	1.12	1.15
	(11)	0.59	0.47	0.52	0.46	0.60	0.64	0.69	0.68
	(9)	0.61	0.48	0.53	0.47	0.66	0.72	0.74	0.71
GLM_a	(10)	1.53	0.76	0.46	0.44	0.25	0.22	0.21	0.23
	(11)	1.68	0.84	0.51	0.50	0.13	0.14	0.12	0.11
	(9)	1.68	0.84	0.52	0.50	0.14	0.14	0.12	0.11
GLM_b	(10)	0.94	0.43	0.29	0.26	0.29	0.33	0.35	0.40
	(11)	0.96	0.42	0.30	0.26	0.23	0.25	0.29	0.30
	(9)	0.97	0.43	0.30	0.27	0.24	0.25	0.30	0.30

pool of candidate models serves as the first critical component of the framework. The pool of candidate models should include those that can be justified by economic considerations and those that are statistically meaningful. Potential deflators should include important independent variables and/or their square roots. In creating such a pool, GLMs should also be considered to handle the scenario where the scale factor is associated with more than one independent variable. Criteria for deflator and model selection mark the second critical component. We argue and show that the two criteria, the average absolute values of studentized residuals and the relative total prediction error for all firm size groups, are valid and effective measures for the scale effect defined as the intriguing combination of coefficient bias and heteroscedasticity. The technique of stratification by a size variable for the assessment of candidate models was also previously used by Easton and Sommers (2003).

Several interesting findings come out from our simulation and sensitivity studies, which may have implications for practical applications. First, if the true scale factor is used as deflator to produce one of the candidate models, this model can be correctly identified using the proposed strategy. Second, if values of A_k and R_k display a nonuniform pattern over the firm groups, the extreme values are always shown up in the first and last few groups. Third, if the total sample size is large, a refined stratification of firms (i.e. more groups) will make nonuniform patterns of values of A_k and R_k more pronounced and therefore make bad models look even worse without causing substantial changes of the uniform pattern for the good ones. Finally and more importantly, the GLM methodology exhibits a great power for mitigating the 'scale effect' under certain scenarios and can be a very useful tool for market-based regression analyses. In the simplest application, GLM can be used to address the estimation problem in commonly encountered scenarios in the level-based regression analysis where larger responses also have larger variations. This approach is statistically more attractive and theoretically sounder than deflation by the response

variable itself. Comparisons between the GLM method and other popular approaches in existing literature, such as White (1980) and MacKinnon and White (1985), require intensive work and are currently under study.

Acknowledgements

This research is partially supported by a grant from the Natural Sciences and Engineering Research Council of Canada. An early version of this article was presented at 2005 AAA Conference in San Francisco, and helpful comments from the discussants are delightfully acknowledged.

References

Akbar, S. and Stark, A. W. (2003) Discussion of scale and the scale effect in market-based accounting research, *Journal of Business Finance and Accounting*, **30**, 57–72.

Barth, M. and Clinch, G. (2001) Scale effects in capital markets-based accounting research, Unpublished Working Paper, Stanford University.

Barth, M. and Kallapur, S. (1996) The effects of cross-sectional scale differences on regression results in empirical accounting research, *Contemporary Accounting Research*, **13**, 527–67.

Christie, A. (1987) On cross-sectional analysis in accounting research, *Journal of Accounting and Economics*, **9**, 233–58.

Easton, P. (1998) Discussion of revalued financial, tangible and intangible assets: association with share prices and non-market-based value estimates, *Journal of Accounting Research*, **36**, 235–47.

Easton, P. and Sommers, G. (2003) Scale and scale effect in market-based accounting research, *Journal of Business Finance & Accounting*, **30**, 25–55.

Lo, K. and Lys, T. (2000) The Ohlson model: contribution to valuation theory, limitations, and empirical applications, *Journal of Accounting Auditing and Finance*, **15**, 337–67.

MacKinnon, J. G. and White, H. (1985) Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties, *Journal of Econometrics*, **29**, 305–25.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, Chapman & Hall, London.

R Development Core Team (2005) R: a language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. Available at http://www.R-project.org

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica*, **48**, 817–38.

Appendix: A Brief Overview of GLM Estimation Theory

In this appendix we provide a short summary of GLMs and the related estimation theory, with particular reference to models useful for market-based regression analyses. More details can be found in the classical reference on GLM by McCullagh and Nelder (1989). Let $(Y_i, X_{1i}, \ldots, X_{ki})$ be the variables of interest and let $\mu_i = E(Y_i)$ be the mean value of the dependent variable. The first major feature of GLM is that the variance of Y_i depends on the mean μ_i through the so-called variance function $V(\cdot)$,

$$\mathrm{Var}(Y_i) = V(\mu_i)\sigma^2$$

and the form of $V(\cdot)$ is assumed to be known. The second major feature of GLM is that the response variable Y_i can be related to $(X_{1i}, \ldots, X_{ki})$ using the mean value μ_i and a link function, $g(\cdot)$.

$$g(\mu_i) = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$$

Once again, the form of $g(\cdot)$ is known. For the classical regression model where $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + e_i$ and $\mathrm{Var}(e_i) = \sigma^2$, it corresponds to a GLM with $g(\mu_i) = \mu_i$ and $V(\mu_i) = 1$. Other forms of variance functions and nonlinear link functions are also permitted. Among popular link functions and variance functions, the identity link $g(\mu_i) = \mu_i$ and the log link $g(\mu_i) = \log(\mu_i)$, the Poisson variance $V(\mu_i) = \mu_i$ and the Gamma variance $V(\mu_i) = \mu_i^2$ might be of particular interest for market-based regression analyses.

The GLM method is semi-parametric and requires specifications only on the first- and second-order moments, i.e. $\mu_i = E(Y_i)$ and $\mathrm{Var}(Y_i)$. The link function specifies how the mean μ_i is related to independent variables and the variance function describes how the variation in response is related to the mean. The combination $\eta_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$ is called the linear predictor while the true mean μ_i might depend on η_i through a nonlinear function, $g(\mu_i) = \eta_i$. The model coefficients $(\beta_0, \beta_1, \ldots, \beta_k)$ are estimated using the maximum quasi-likelihood method. For the most general case, the estimator, denoted by $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)'$, is the solution to the following quasi-score equation:

$$\mathbf{D}'\mathbf{V}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{0} \qquad \text{(A1)}$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$, $\mathbf{V} = \text{diag}(V(\mu_1), \ldots, V(\mu_n))$ and $\mathbf{D} = \partial\mu/\partial\beta$. The model parameters β_i are hidden inside (A1) through $\boldsymbol{\mu}$, $\mathbf{V}$ and $\mathbf{D}$. In the simple case of linear model where $\mu_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki}$, we have $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ and $\mathbf{D} = \mathbf{X}$, where $\mathbf{X}$ is the usual design matrix for the regression model. The quasi-score Equation A1 becomes $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$, which is equivalent to the normal equation used for the WLS estimation if we treat the variance matrix $\mathbf{V}$ as the weight matrix. If $V(\mu_i) = v_i$ is a known constant, then the solution to (A1) is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}')^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$, the WLS estimator. But if $V(\mu_i)$ depends on μ_i, then the weight matrix $\mathbf{V}$ involves the regression parameters β, and consequently no closed form solution exists for (A1).

The major difficulty in finding the maximum quasi-likelihood estimator $\hat{\beta}$ is that one typically needs to solve (A1) using iterative procedures. The statistical softwares S-PLUS and R, among others, have built-in functions for fitting GLMs. The R package works almost the same as S-PLUS but is free for research use and downloadable from the R-project homepage. Suppose we wish to find the maximum quasi-likelihood estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ for the model $\mu_i = E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$ and $\text{Var}(Y_i) = \boldsymbol{\mu}_i^2\sigma^2$ with Y, X1 and X2 being the vectors of sample data for each of the variables. The following lines show how to obtain $\hat{\boldsymbol{\beta}}$ using R or S-PLUS.

```
b0<−lm(Y ~ X1 + X2)$coefficients
b1<−glm(Y ~ X1 + X2, start = b0,
quasi(var = ''mu^2'',link =
''identity''))$coefficients
```

The b0 is the OLS estimator which serves as the initial value for the GLM estimator b1. The function glm usually requires that all Y_i's be positive. Other variance function options include var = 'mu' and var = 'constant', and the link function could for instance be link = 'log'. Another way to find $\hat{\boldsymbol{\beta}}$ is to write a specific R program for each of the models under consideration. This is what we used in this study. The following R/S-PLUS program computes $\hat{\boldsymbol{\beta}}$ for $g(\mu_i) = \mu_i$ and $V(\mu_i) = \mu_i^2$, with Y, X1 and X2 being the vectors of sample data for each of the variables as in the previous example. Let n be the sample size.

```
tol<−1e-08
dif<−1
int<−rep(1,n)
X<−cbind(int,X1,X2)
b0<−solve(t(X)%*%X,t(X)%*%Y)
while(dif > tol){
mu<−as.vector(X%*%b0)
XX<−cbind(int/mu^2,X1/mu^2,X2/mu^2)
D<−solve(t(X)%*%XX,t(XX)%*%(Y−mu))
dif<−max(abs(D))
b1<−b0 + 0.1*D
b0<−b1}
```

If the variance function is $V(\mu_i) = \mu_i$, one needs to modify the line involving XX using `XX< − cbind(int/mu, X1/mu, X2/mu)`. The fitted (or predicted) value for Y_i at $x_i = (1, X_{1i}, \ldots, X_{ki})'$ is computed as $\hat{\mu}_i = g^{-1}(x_i'\hat{\boldsymbol{\beta}})$, where $g^{-1}(\cdot)$ is the inverse of the link function; the method of moment estimator for σ^2 is given by $\hat{\sigma}^2 = (n-k-1)^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)^2/V(\hat{\mu}_i)$; and the studentized residuals are computed as $r_i = (Y_i - \hat{\mu}_i)/(\hat{\sigma}\sqrt{V(\hat{\mu}_i)})$.