A Practical Suggestion for Log-Linear Workers Compensation Cost Models

Dan Corro
December, 1998

Abstract

It is standard practice to use log-linear and log-log regression models in the analysis of workers compensation claim costs. While useful for the investigation of proportional cost relationships, those transformed models are not well suited for predicting individual or even average claim costs. There is, however, enormous potential for using regression equations as a computational device for generating tabular reserves and for benchmarking select sets of claim costs. This paper suggests that changing the assigned weights of observations in the determination of the logged cost model can improve its predicted values for conversion back to a dollar scale. The derivation of a specific reweighting formula is motivated from the basic data fitting geometry of OLS regression. The technique is tested via a simulation and on a large database of actual workers compensation lost time claims. Both show a marked improvement in the claim cost estimates determined from the modified regression equations.

Introduction

The use of log-linear and or log-log regression models is the preferred practice for the analysis of workers compensation insurance claim costs. The use of a logarithmic scale generally renders the cost distribution pattern more symmetric and less influenced by large "outlier" claims. It has the additional advantage of not predicting negative costs. Some of the more formal advantages are discussed in the background section below. While this typically results in better fits and higher R^2 values, it is well known that the attempt to reverse the transformation by exponentiation usually fails to yield very useful dollar cost estimates. Indeed, on average the figures which result are smaller--sometimes spectacularly smaller-- than the original costs used to construct the model. As explained in the paper, this is a formal consequence of the geometric mean cost being less than the arithmetic mean. While the transformed models provide useful information on cost relationships, that transformation renders them of little value for directly predicting dollar cost estimates.

The common sense explanation for this is that the high cost claims are effectively given less weight in a log-linear model. This is viewed as one of the prices to be paid for mitigating the influence of outlier claims. This paper pursues this somewhat further but from a simple geometric point of view rather than from the more challenging perspective of model specification error. It begins with the observation that cost data is typically presented with a "natural weight". This may simply be one claim one vote within a claim population or, as is often the case, a weight inferred from claim sampling procedures or other information on the probability of claim occurrence. It is key that this "natural" quality in dollar terms need not be preserved under transformation of the data. In particular, this may happen when costs are recalibrated via the log function. This paper illustrates this and suggests that one practical way to deal with it is to reweight the data. Reweighting observations is a common practice in constructing regression models to temper the effect of outliers or more generally to deal with heteroscedasticity. What is presented here is a much more simpleminded application, one designed simply to shift the focal point of a logged cost model so as to make it better suited to producing dollar cost estimates.

The task of generating tabular reserves is a natural application of multivariate analysis, provided the cost models yield suitable estimates when expressed in dollar terms. A single regression equation can be the formal equivalent of innumerable tables. Fitted cost equations have the potential to simultaneously account for many and varied claim characteristics and are computationally very convenient. For the most part, claims research has produced cost models geared toward determining proportional cost effects rather than dollar cost estimates. The ability to tune the regression equation to the task of estimating the average cost per case is therefore of practical importance.

The paper begins with a background section dealing particularly with the treatment of categorical information and claim sampling in this context. The next section illustrates how linear regression focuses on the "center of gravity" and shows how alternative weighting schemes may be used to shift that focus. The following section presents some simple numeric facts and establishes the existence and essential uniqueness of an "exponential adjusted weight" which is

shown, in the following section, to possess its own "naturalness" in the context of logged cost estimation. The next two sections give examples. The first is simulation based and confirms some expected properties of alternative weights. The second example is of particular importance as it details the results of applying this technique to a large workers compensation lost time claim data base. The final section draws some conclusions and suggests an avenue of further study. An Appendix provides a SAS routine for determining the exponential adjusted weight, which the paper puts forth as a useful tool in the application of logged cost models.

While much of this paper is devoted to properties of the exponential adjusted weight, this is not meant to imply that this is **the** correct choice for use with log-linear regression models. Indeed, the view is that weights correspond to perspectives on the data and no one perspective is best for this purpose. Additional perspectives--some via alternative weights--can provide insights for understanding the cost model. The work illustrates how transforming the data may dictate the need to consider such additional perspectives.

Background

Much of the information captured in workers compensation claim databases is "categorical" data, that is, it categorizes claims. For example, many databases include categories which identify the part of body injured or the cause of the injury. Some databases capture administrative categories, such as controverted cases or those for which the claimant has sought legal representation. Such categorical values are often incorporated into models designed to explain or predict claim costs. Usually this is in combination with numeric claim data, such as claimant age or pre-injury wage.

One of the most robust sources of workers compensation claim information is the Detailed Claim Information [DCI] database of the National Council on Compensation Insurance, Inc. Introduced in 1979 and expanded in 1991 to include 42 jurisdictions, the DCI is a stratified random sample of lost time cases. The DCI captures over 80 data elements on over 700,000 claims. For claims selected for the sample, an initial DCI report is required 6 months after notification to the insurer with annual follow-up reports through claim closure (up to 9 reports). The categorical data elements include: the state of jurisdiction, the report sequence number and resolution status of the claim as of that report, the claimant's gender and marital status, the part of body injured together with the nature and cause of the injury, whether the claimant was represented by an attorney and whether surgery or hospitalization was required, etc. The numeric data elements captured in the DCI include: the date of the injury and where applicable the date of return to work, the age and wage of the claimant at the time of injury, an itemization of payments and case reserve for medical care and compensation for lost time, and loss adjustment costs allocated to the claim, etc. An empirical example is presented below which illustrates the findings using DCI data.

Let x represent an observation, $z = z_x$ the corresponding claim cost and $\{x_i\}$ the values of a set of explanatory variables. This note considers log cost models of the form:

$$y = \ln(z) = \Sigma \beta_i x_i + u$$

where u represents the error term. The x_i may be categorical or continuous and, if continuous, in its original scale (log-linear cost model) or transformed to a logarithmic scale (log-log cost model).

On the continuous side, pre-injury wage and rate of compensation are important examples. Typically, dollar amounts like the pre-injury wage would be logged while that need not be the case for other continuous variables, such as the rate of compensation (periodic lost time compensation expressed as a percentage of the wage). Observe that the model parameter β_i does not vary with claim cost z, referred to as an assumption of constant elasticity (for x_i in logged form). It is common to use the full wage (or log thereof) so as to capture utilization effects related with total income. This is done even though workers compensation benefit statutes impose maximum wage replacement levels. Their presence, it can be argued, compromises the assumption of constant elasticity. In any event, it is worth considering the implications on the use of the regression equation when $\{\beta_i\}$ is observed to vary with z.

The appeal of a log cost model in this context is, however, most easily recognized in the case of categorical variables. In the simplest case, suppose that the explanatory variable x_i corresponds to a {yes,no} condition, taking on the respective values {1,0}. In terms of the original cost z, the model associates an adjustment factor of $\alpha_i = e^{\beta_i}$. Most categories are better associated with such a proportional shift than to a particular dollar amount, as would occur if the logarithm were not used to transform the dependent variable of the cost model. While the above remarks on the shape of loss distributions is an important consideration, it is this observation together with the desire to avoid negative cost estimates which provides the strongest motivation for using logarithms to model workers compensation claim costs.

As with continuous variables, there is the issue as to whether the adjustment factor α_i associated with a characteristic variable changes with z. Consider, for example, the characteristic whether a claimant is represented by an attorney. For most purposes it is clearly preferable to model the associated cost impact as a proportional rather than as a flat loading. Again there are countervailing statutory benefit considerations: some state statutes regulate attorney fees by imposing maximums or sliding scales relative to the settlement amount.

The expense of collecting and storing detailed information on every claim may be prohibitively high, so oftentimes cost analyses resort to using claim samples. The efficiency of the claim sampling process may be further improved through stratification. In the case of the DCI database, state specific sampling ratios are used and stratification is applied so that the relatively simple and quickly resolved cases--for which many of the claim characteristics are missing or inapplicable--do not bog down the collection, storage and processing tasks. In this situation, a weight variable would be applied in deriving a cost model. In this study we use the notation $\omega_x (= \omega_{y_x} = \omega_{z_x})$ to denote the weight assigned to the claim x based upon the sampling rules. In the case of the DCI, ω_x is determined as the inverse of the applicable state sampling ratio, selectively increased by a factor to account for stratification. The set of weights, $\{\omega_x\}$, have the very desirable feature that, assuming the sampling is done correctly, the corresponding weighted arithmetic mean is an unbiased estimator of the average cost per case of lost time claims. Although not necessarily an integer, the value ω_x can be interpreted as the number of claims represented by the sampled claim x. When the set $\{\omega_x\}$ is this sampling weight, the sum total $W = \Sigma \omega_x$ provides an estimate of the size of the lost time claim population. Making the normalization $p_x = \frac{\omega_x}{W}$ converts the weights into a probability density and reveals the weighted mean to equal the expected claim cost:

$$E(z) = \Sigma p_x z_x = \frac{\Sigma \omega_x z_x}{\Sigma \omega_x}.$$

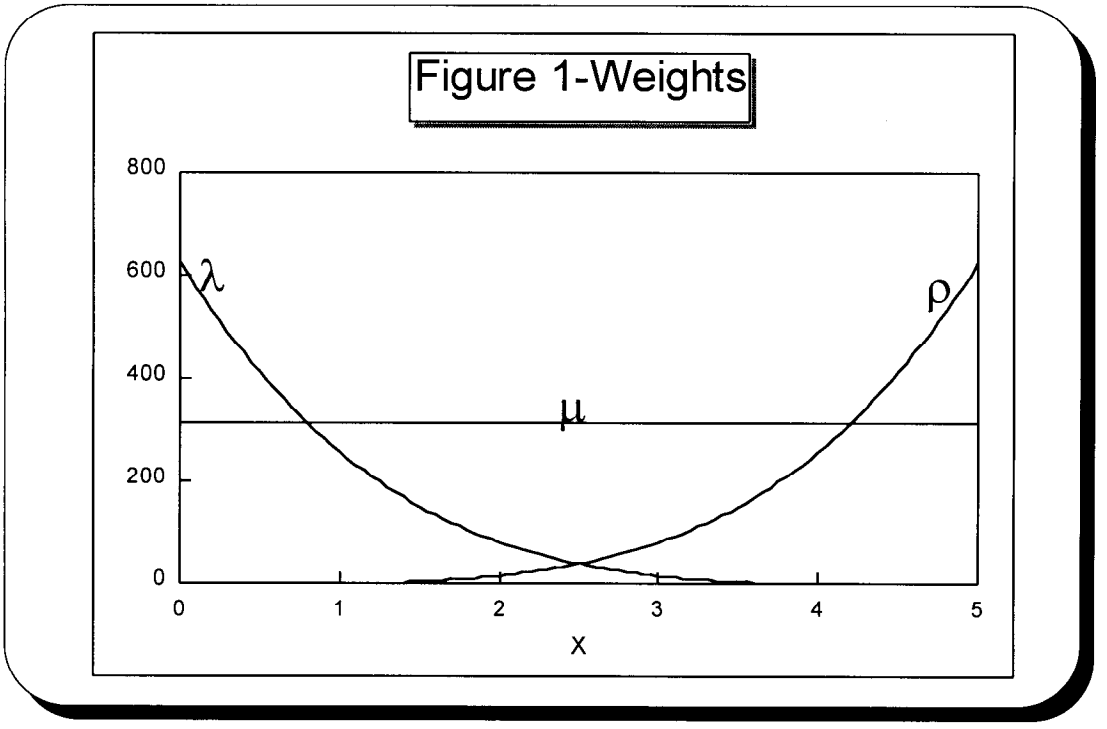
Weights and Linear Regression--a Geometric Example

The object of this section is to illustrate in very simple geometric terms how a choice of weight effects the regression equation and, by implication, how the intended application of a regression equation may suggest a choice of weight. This is done via a simple example: consider the set of points along the quadrant of a circle:

$$y = f(x) = \sqrt{x(10-x)} \text{ for } x \in \{0, 0.05, 0.10, \ldots, 5.00\}$$

This defines the dependent variable y from a deterministic relationship with the single explanatory variable x. Observe that while y increases with x, its rate of change depends on its value.

Three weights are considered, as depicted in Figure 1:



Weight λ: skewed toward left (smaller y)

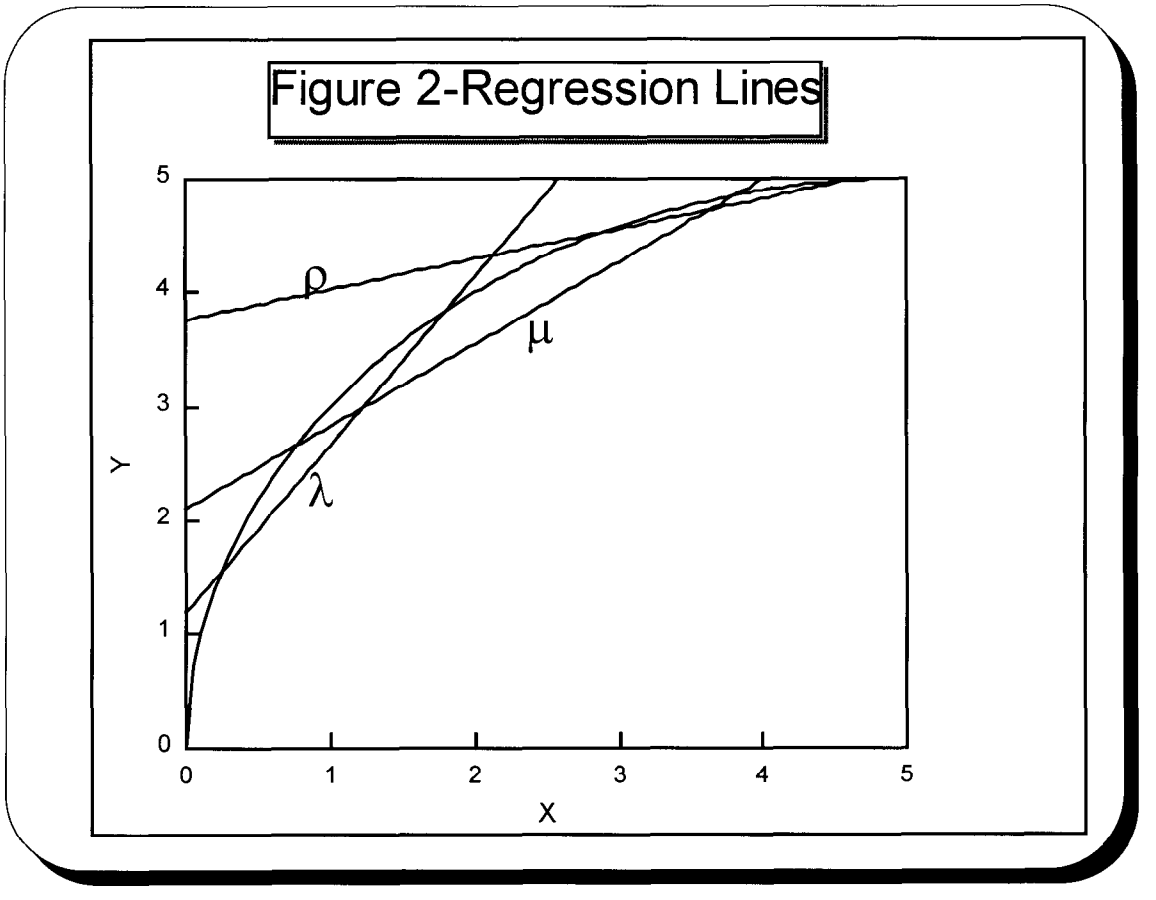
Weight μ: uniform (independent of y)

Weight ρ: skewed toward right (larger y)

For each weight, the regression equation

$$\hat{y} = \alpha + \beta x$$

is plotted in Figure 2, along with the circle observation points:



Recall that the center of gravity $(\bar{x}, \bar{y})$ lies on the regression equation. The following table shows how this relates with the choice of weight:

Weight	$\bar{x}$	$\bar{y}$	β	$f'(\bar{y})$
λ	0.81	2.39	1.48	1.84
μ	2.50	3.91	0.72	0.80
ρ	4.19	4.88	0.27	0.22

The regression equation behaves as would be expected when the weight shifts from left to right. This simple example illustrates how the choice of weight provides a means to customize the "slope" parameter β to a specified value of $y \approx \bar{y}$.

Along with focusing the slope parameter to the center of gravity, OLS estimation is optimized at that point (i.e. for any fixed confidence level, the length of the interval about the point estimate determined from the regression equation is minimized at $x = \bar{x}$--this applies whether the estimate is of a single observation or of an average). It follows that if a particular y value is of interest, an appropriate choice of weight may render the regression equation more suitable. If, for example, the largest values $y \approx 5$ were of greatest concern, ρ is the preferable of the three weights.

The observations of this section are formal consequences of the data fitting properties of the OLS regression equation. It is independent of how well or poorly the linear model may be for the purpose of statistical estimation. Indeed, these remarks are of least interest when a linear model is well specified, since the choice of weight would then be expected to have a negligible effect upon the model parameter β .

Numeric Relationships

It is well known that the geometric mean is less than the arithmetic mean:

$$[\Pi_{i=1}^{N} y_i]^{\frac{1}{N}} \leq \frac{1}{N}\Sigma_{i=1}^{N} y_i$$

for any set of N nonnegative real numbers $\{y_i\}$. This section generalizes the above inequality and establishes the existence and essential uniqueness of a modified (reweighted) geometric mean which is useful in relating proportional effects with arithmetic means. The following lemma establishes the requisite inequalities. It is basically a first semester calculus exercise.

Lemma: For $\lambda \in (0, 1]$ and $\tau \in [0, 1]$

a) $\tau + \lambda - \tau\lambda - \lambda^{1-\tau} \geq 0$

b) $\lambda^{\frac{(1-\tau)\lambda}{\tau+(1-\tau)\lambda}} - \tau - \lambda + \lambda\tau \geq 0$

Proof: a) Fix τ and define

$$f(\lambda) = \tau + \lambda - \tau\lambda - \lambda^{1-\tau}$$

and note that

$$\frac{df}{d\lambda} = 1 - \tau - (1-\tau)\lambda^{-\tau} = (1-\tau)(1-\lambda^{-\tau}) \leq 0$$

because $\tau \leq 1$ and $\lambda^{-\tau} \geq 1$. This implies that $f(\lambda)$ is decreasing on (0,1] and so

$$\tau + \lambda - \tau\lambda - \lambda^{1-\tau} = f(\lambda) \geq f(1) = \tau + 1 - \tau - 1^{1-\tau} = 0$$

establishing a). Note that the argument implies strict inequality for $\lambda, \tau \in (0, 1)$.

b) Again fix τ , define

$$g(\lambda) = \frac{(1-\tau)\lambda}{\tau+(1-\tau)\lambda}$$

and note that $0 \leq g(\lambda) \leq 1$. Also define

$$h(\lambda) = g(\lambda)\lambda^{g(\lambda)-1}$$

Logarithmic differentiation of $g(\lambda)$ implies that

$$\frac{dg}{d\lambda} = \frac{g(\lambda)(1-g(\lambda))}{\lambda} \geq 0$$

Logarithmic differentiation of $h(\lambda)$ implies that

$$\frac{dh}{d\lambda} = \frac{\log(\lambda)h(\lambda)g(\lambda)(1-g(\lambda))}{\lambda} \leq 0$$

since $\lambda \leq 1 \Rightarrow \log(\lambda) \leq 0$ and all the other factors are ≥ 0. Since $g(\lambda)$ is increasing and $h(\lambda)$ is decreasing:

$$h(\lambda) \geq h(1) = g(1)1^{g(1)-1} = 1 - \tau = g(1) \geq g(\lambda)$$

Consider the function

$$f(\lambda) = \lambda^{g(\lambda)} - \tau - \lambda + \lambda\tau$$

Logarithmic differentiation of the term $\lambda^{g(\lambda)}$ combined with the above expression for $\frac{dg}{d\lambda}$ and observation that $\log(\lambda) \leq 0$ gives:

$$\frac{df}{d\lambda} = g(\lambda)\lambda^{g(\lambda)-1} + \log(\lambda)[\lambda^{g(\lambda)-1}g(\lambda)(1-g(\lambda))] - 1 + \tau$$

$$\geq g(\lambda)\lambda^{g(\lambda)-1} - 1 + \tau = h(\lambda) - 1 + \tau \geq (1-\tau) - 1 + \tau = 0$$

It follows that $f(\lambda)$ is increasing, whence:

$$\lambda^{\frac{(1-\tau)\lambda}{\tau+(1-\tau)\lambda}} - \tau - \lambda + \lambda\tau = f(\lambda) \leq f(1) = 1 - \tau - 1 + \tau = 0$$

which establishes b) and completes the proof of the lemma.

Let $\Re^+$ denote the positive real numbers, the following generalizes the relationship between the geometric and arithmetic averages to "weighted" averages.

Propositon 1: For any $\{y_i, \omega_i | 1 \le i \le N\} \subset \Re^+$,

$$[\Pi_{i=1}^N y_i^{\omega_i}]^{\frac{1}{W}} \le \frac{1}{W}\Sigma_{i=1}^N \omega_i y_i \qquad \text{where } W = \Sigma_{i=1}^N \omega_i$$

Proof: Induction on N. The case $N = 1$ is clear $[(y_1^{\omega_1})^{\frac{1}{\omega_1}} = y_1 = \frac{\omega_1}{\omega_1} y_1]$. It is instructive to do the case $N = 2$. By symmetry, we may assume that $y_1 \ge y_2$ and replacing ω_i with $\frac{\omega_i}{W}$, we may assume that $W = 1$. Part a) of the lemma, with $\lambda = \frac{y_2}{y_1}$ and $\tau = \omega_1$ gives:

$$\omega_1 y_1 + \omega_2 y_2 = [\tau + \lambda - \lambda\tau] y_1 \ge \lambda^{1-\tau} y_1 = y_1^{\tau}(\lambda y_1)^{1-\tau} = y_1^{\omega_1} y_2^{\omega_2}$$

which establishes the case $N = 2$. Now proceed by induction. As before, we may assume without loss of generality that $W = 1$. Letting $\hat{W} = 1 - \omega_N$, the induction hypothesis gives:

$$[\Pi_{i=1}^{N-1} y_i^{\omega_i}]^{\frac{1}{\hat{W}}} \le \frac{1}{\hat{W}}\Sigma_{i=1}^{N-1} \omega_i y_i$$

Letting $z_1 = \frac{1}{\hat{W}}\Sigma_{i=1}^{N-1} y_i$ and $z_2 = y_N$, the case $N = 2$ and the above inequality combine to give

$$\Sigma_{i=1}^N \omega_i y_i = \hat{W} z_1 + \omega_N z_2 \ge z_1^{\hat{W}} z_2^{\omega_N} \ge [[\Pi_{i=1}^{N-1} y_i^{\omega_i}]^{\frac{1}{\hat{W}}}]^{\hat{W}} y_N^{\omega_N}$$
$$= \Pi_{i=1}^N y_i^{\omega_i}$$

which completes the proof.

The fact that the weighted geometric mean is always less than the weighted arithmetic mean suggests that an alternative weight--one shifting more weight to larger values--may offset this. The simplest alternative is a weight proportional to the value itself. The following proposition shows that choice overcompensates:

Proposition 2: For all $\{y_i, \omega_i | 1 \le i \le N\} \subset \Re^+$

$$[\Pi_{i=1}^N y_i^{\omega_i y_i}]^{\frac{1}{V}} \ge \frac{1}{W}\Sigma_{i=1}^N \omega_i y_i \text{ where } W = \Sigma_{i=1}^N \omega_i,\ V = \Sigma_{i=1}^N \omega_i y_i$$

Proof: The proof is again by induction. The case $N = 1$ is clear $[y_1^{\frac{\omega_1 y_1}{\omega_1 y_1}} = y_1 = \frac{\omega_1}{\omega_1} y_1]$. It is again helpful to do the case $N = 2$. By symmetry, we may assume that $y_1 \ge y_2$ and, replacing ω_i with $\frac{\omega_i}{W}$, that $W = 1$. Again letting $\lambda = \frac{y_2}{y_1}$ and $\tau = \omega_1$, part b) of the lemma gives

$$\omega_1 y_1 + \omega_2 y_2 = [\tau + (1-\tau)\lambda] y_1 \le [\lambda^{\frac{(1-\tau)\lambda}{\tau+(1-\tau)\lambda}}] y_1$$

$$= y_1^{\frac{\tau}{\tau+(1-\tau)\lambda}} [\lambda y_1]^{\frac{(1-\tau)\lambda}{\tau+(1-\tau)\lambda}} = y_1^{\frac{\tau y_1}{V}} y_2^{\frac{(1-\tau)\lambda y_1}{V}} = [y_1^{\omega_1 y_1} y_2^{\omega_2 y_2}]^{\frac{1}{V}}$$

which establishes the case $N=2$. Proceed by induction. As usual, we may assume $W=1$. Let $\hat{W}=1-\omega_N$, $\hat{V}=V-\omega_N y_N$, the induction hypothesis gives:

$$[\Pi_{i=1}^{N-1} y_i]^{\frac{1}{\hat{V}}} \geq \frac{1}{\hat{W}} \Sigma_{i=1}^{N-1} \omega_i y_i = \frac{\hat{V}}{\hat{W}}$$

Let $z_1 = \frac{\hat{V}}{\hat{W}}, z_2 = y_N$ and note that: $z_1 \hat{W} + \omega_N z_2 = \hat{V} + \omega_N y_N = V$. The above inequality and the case $N=2$ give:

$$[\Pi_{i=1}^{N} y_i^{\omega_i y_i}]^{\frac{1}{V}} = [[\Pi_{i=1}^{N-1} y_i]^{\frac{1}{\hat{V}}}]^{\frac{\hat{V}}{V}} y_N^{\frac{\omega_N y_N}{V}} \geq [\frac{\hat{V}}{\hat{W}}]^{\frac{\hat{V}}{V}} y_N^{\frac{V-\hat{V}}{V}}$$

$$= z_1^{\frac{\hat{W} z_1}{V}} z_2^{\frac{\omega_N z_2}{V}} \geq \hat{W} z_1 + \omega_N z_2 = V = \frac{1}{W} \Sigma_{i=1}^{N} \omega_i y_i$$

completing the proof.

It follows that letting $\mu = \frac{1}{W} \Sigma \omega_i y_i$ we have:

$$[\Pi y_i^{\omega_i}]^{\frac{1}{\Sigma \omega_i}} \leq \mu \leq [\Pi y_i^{\omega_i y_i}]^{\frac{1}{\Sigma \omega_i y_i}}$$

It is natural, then, to consider weights $\{\gamma_i\}$ satisfying:

$$\mu = [\Pi y_i^{\gamma_i}]^{\frac{1}{\Sigma \gamma_i}}$$

It turns out that such equations determine an essentially unique set of weights. This is a corollary of the following:

Proposition 3: For any $F:\Re^+\to\Re^+$ which is monotonic increasing and any $\{z_i,\omega_i | z_i\le z_{i+1}, 1\le i\le N\}\subset\Re^+$ there is a set $\{\gamma_i | 1\le i\le N\}$, uniquely determined up to a positive scaler multiple, such that

$$\frac{\sum_{i=1}^{k}\gamma_i F(z_i)}{\sum_{i=1}^{k}\gamma_i}=F\left(\frac{\sum_{i=1}^{k}\omega_i z_i}{\sum_{i=1}^{k}\omega_i}\right),\quad 1\le k\le N.$$

Proof: Induction on N, the case $N=1$ being trivial. We may assume, without loss of generality, that $W=\Sigma_{i=1}^{N}\omega_i=1, \omega_N>0$ and that $z_i<z_{i+1}, 1\le i\le N-1$. Let

$$y_2=\Sigma_{i=1}^{N}\omega_i z_i,\quad y_1=\frac{\sum_{i=1}^{N-1}\omega_i z_i}{\sum_{i=1}^{N-1}\omega_i},$$

then observe that $y_1\le z_{N-1}<z_N$ and so:

$$y_1=(1-\omega_N)y_1+\omega_N y_1<(1-\omega_N)y_1+\omega_N z_N=y_2\le z_N.$$

Since F is monotonic increasing, it follows that

$$F(y_1)<F(y_2)\le F(z_N)$$

which enables us to define

$$\alpha=\frac{F(z_N)-F(y_2)}{F(z_N)-F(y_1)}\in[0,1).$$

By induction, there is a uniquely determined set $\{\gamma_i | 1\le i\le N-1\}$ such that

$$\frac{\sum_{i=1}^{k}\gamma_i F(z_i)}{\sum_{i=1}^{k}\gamma_i}=F\left(\frac{\sum_{i=1}^{k}\omega_i z_i}{\sum_{i=1}^{k}\omega_i}\right),\quad 1\le k\le N-1,\ \Sigma_{i=1}^{N-1}\gamma_i=\alpha$$

Setting $\gamma_N=1-\alpha\in\Re^+$, we have $\Sigma_{i=1}^{N}\gamma_i=1=\Sigma_{i=1}^{N}\omega_i$ and

$$\Sigma_{i=1}^{N}\gamma_i F(z_i)=\Sigma_{i=1}^{N-1}\gamma_i F(z_i)+(1-\alpha)F(z_N)$$

$$=\alpha F\left(\frac{\sum_{i=1}^{N-1}\omega_i z_i}{\sum_{i=1}^{N-1}\omega_i}\right)+(1-\alpha)F(z_N)=\alpha F(y_1)+(1-\alpha)F(z_N)$$

$$= F(z_N) - \alpha[F(z_N) - F(y_1)] = F(z_N) - [F(z_N) - F(y_2)]$$

$$= F(y_2) = F(\Sigma_{i=1}^{N} \omega_i z_i)$$

Observe that we in effect solved for

$$\gamma_N = \frac{F(y_2)-F(y_1)}{F(z_N)-F(y_1)}$$

which establishes uniqueness and completes the proof.

Corollary: For any $\{z_i, \omega_i | 1 \le i \le N\} \subset \Re^+$, there is a unique set $\{\gamma_i | 1 \le i \le N\} \subset \Re^+$ such that

$$[\Pi_{i=1}^{k} z_i^{\gamma_i}]^{\frac{1}{\Sigma_{i=1}^{k} \gamma_i}} = \frac{\Sigma_{i=1}^{k} \omega_i z_i}{\Sigma_{i=1}^{k} \omega_i}, 1 \le k \le N \text{ and } \Sigma_{i=1}^{N} \gamma_i = \Sigma_{i=1}^{N} \omega_i$$

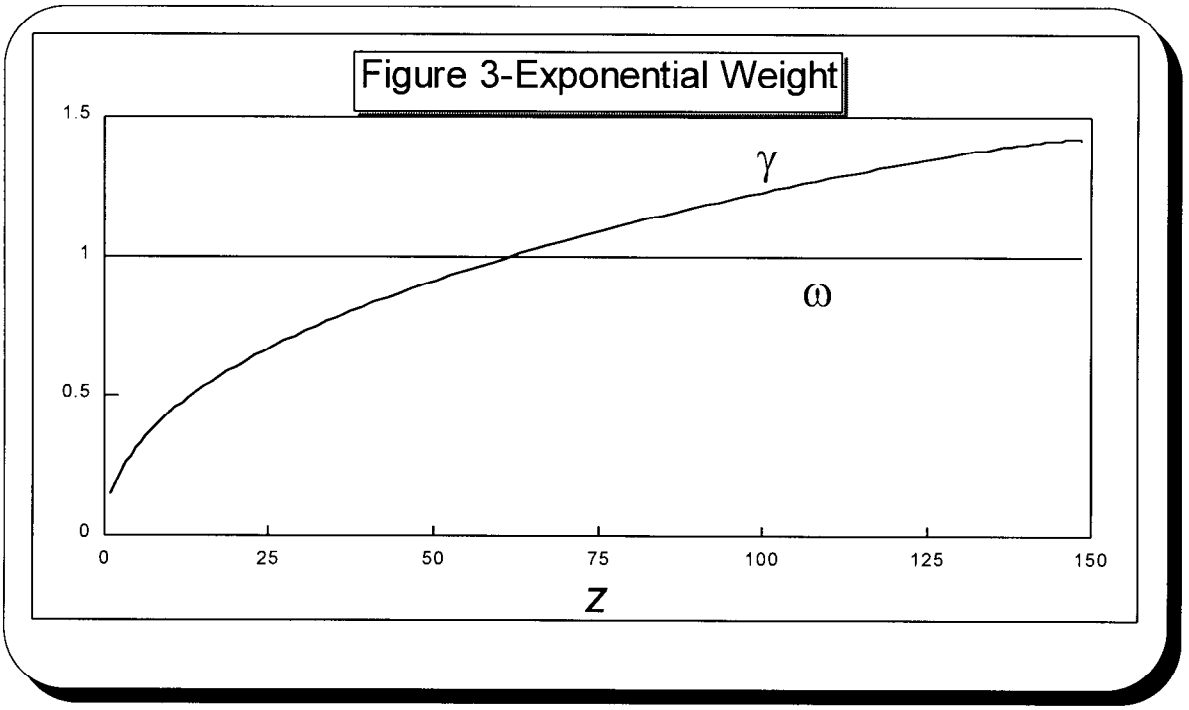
Proof: Letting $F(z) = \log(z)$ in Proposition 3, there is a unique set $\{\gamma_i | 1 \le i \le N\} \subset \Re^+$ such that:

$$\Sigma_{i=1}^{N} \gamma_i = \Sigma_{i=1}^{N} \omega_i, \quad \frac{\Sigma_{i=1}^{k} \gamma_i \log(z_i)}{\Sigma_{i=1}^{k} \gamma_i} = \log\left(\frac{\Sigma_{i=1}^{k} \omega_i z_i}{\Sigma_{i=1}^{k} \omega_i}\right), 1 \le k \le N$$

and the result follows by exponentiation.

The set $\{\gamma_i\}$ can be interpreted as a reweighting of $\{z_i, \omega_i\}$ with the same total weight and which makes the reweighted geometric mean equal the original weighted arithmetic mean. This can be viewed as a weight shift designed to offset the effect of the logarithmic transformation and is accordingly referred to as the *exponential adjusted weight*.

Letting $z = e^y$, $\omega \equiv 1$ in the circle example of the previous section, Figure 3 shows the exponential weight γ as a function of z.



Remark 1: The last line of the proof of Proposition 1 provides a recursive formula which can in practice be used to calculate the γ_i (refer to the Appendix).

Remark 2: The potential degrees of freedom in the definition of a weight variable is large ($\approx N$, $N-1$ if the total weight is specified). This was exploited in the derivation of the exponential weight γ so as to impose constraints over an expanding series of cost intervals. It has been observed that the two geometric means determined using the sampling weight ω and using the proportional cost weight ρ are respectively less than and greater than the arithmetic mean cost μ. It is therefore natural to consider the 2-dimensional family of weights of the form:

$$\eta = a\omega + b\rho, \quad a, b \in \Re^+$$

Observe that the requirement $\Sigma\eta = \Sigma\rho = \Sigma\omega = W$ removes a degree of freedom and forces $b = 1 - a$. This leaves 1 degree of freedom and suggests a closed form solution. Indeed, set

$$\mu = \frac{\sum \omega z}{W}, \ \lambda = [\prod z^{\omega}]^{\frac{1}{W}}, \ \kappa = [\prod z^{\rho}]^{\frac{1}{W}}$$

Assume, to avoid degeneracy, that there are at least two distinct values of z with positive weight $\omega = \omega_z > 0$. It has been shown that $\lambda < \mu \leq \kappa$ which assures that

$$a = \frac{\log(\mu)-\log(\lambda)}{\log(\kappa)-\log(\lambda)} \in (0,1)$$

is well defined. Setting $\eta = a\omega + (1-a)\rho$, it follows that:

$$\Sigma\eta = \Sigma\omega = W, \qquad [\Pi z^{\eta}]^{\frac{1}{W}} = \mu = \frac{1}{W}\Sigma\omega z$$

This η differs from the exponential adjusted weight γ as defined here. Indeed, in the example illustrated in Figure 3, γ is clearly **not** a linear combination of $\omega \equiv 1$ and $\rho = \frac{z}{W}$.

Properties of the Exponential Weight

Letting a bar denote the weighted mean taken over the claim sample:

$$\bar{u} = 0 \quad \bar{y} = \overline{\ln(z)} = \Sigma \beta_i \bar{x}_i$$

Regressions provide their best fit at this mean. Assume that the original set of weights $\{\omega_x\}$ is used in evaluating the log cost model. In dollar terms, therefore, it follows that the model parameters $\{\beta_i\}$ are optimized not at the original arithmetic mean but at the weighted geometric mean. Indeed, when exponentiation is used to convert the mean logged cost $\bar{y}$ back to dollar terms:

$$\hat{z} = e^{\bar{y}} = e^{\frac{1}{W}\sum \omega_x \ln(z_x)} = [\prod z_x^{\omega_x}]^{\frac{1}{W}}$$

As point estimates, the factors $\alpha_j = e^{\beta_j}$ which the model associates with the explanatory variables are proportional adjustment factors applicable to the geometric mean $\hat{z}$.

It is, however, the simple average cost per case which is the standard measure of claim severity. Geometric means, typically much smaller, do not have as straightforward a relationship with incurred costs. Indeed, claim costs are often decomposed into frequency and severity components with the severity component being a (weighted) arithmetic mean. Since the proportional effects may well vary with z, the dollar size of the claim, it may be desirable to refocus the regression model to the higher arithmetic cost level. This paper proposes the use of alternative sets of weights--the exponential weight in particular--as a way to achieve this.

Intuitively, the translation to logarithms has the effect of making claims more "equal". In particular, the high cost claims have less influence in the mean. A natural correction to this is a scheme which assigns more weight to higher cost claims when evaluating the regression model. The simplest way to achieve this is to make the weight of an observation proportional to the dollar cost. Of course, the original sampling rules must not be ignored. This motivates the definition of an alternative set of weights $\{\rho_x\}$, referred to here as the proportional cost weight, which for each claim x_0 is defined as:

$$\rho_{x_0} = \frac{\sum \omega_x}{\sum \omega_x z_x} \omega_{x_0} z_x$$

where the sums are over all claims and the ratio is imposed so that the total proportional weight equals the same total weight W as do the original set of sampling weights.

The previous section shows that given any weighting of the original cost variable z there is a related exponential adjusted weight which can be assigned to the values taken on by the transformed dependent variable $y = \log(z)$. The main point is that the use of that alternative weight in determining the regression equation serves to optimize its fit to the data at a point

corresponding to the usual mean cost per case. Indeed, the previous section established the existence and essential uniqueness of this weighting and provides a practical means for its determination. This alternative weight $\{\gamma_x\}$, referred to as the exponential weight, is the weight related in the requisite way with the original sampling weight $\{\omega_x\}$. Again holding the total weight W constant, among the defining properties is:

$$\log\left(\frac{\sum \omega_x z_x}{W}\right) = \frac{\sum \gamma_x \log(z_x)}{W}$$

from which it follows that

$$\bar{z} = \frac{\sum \omega_x z_x}{W} = \exp\left(\frac{\sum \gamma_x \log(z_x)}{W}\right) = \left[\prod z_x^{\gamma_x}\right]^{\frac{1}{W}}$$

The previous section also established the relationships:

$$\left[\prod z_x^{\omega_x}\right]^{\frac{1}{W}} \leq \bar{z} \leq \left[\prod z_x^{\rho_x}\right]^{\frac{1}{W}}$$

which suggest that using the original sampling weight $\{\omega_x\}$ to evaluate the log cost model will produce a mean cost estimate which is biased downward, whereas use of the proportional cost weight $\{\rho_x\}$ will overstate the mean cost.

When the exponential weight $\{\gamma_x\}$ is used to evaluate the log cost model, it follows that:

$$\bar{z} = \exp\left(\frac{\sum \gamma_x y_x}{W}\right) = \exp\left(\frac{\sum\sum \gamma_x \beta_i x_i}{W}\right)$$

which implies that

$$\bar{z} = \prod \alpha_i^{c_i} \text{ where } \alpha_i = e^{\beta_i} \text{ and } c_i = \frac{\sum \gamma_x x_i}{W}.$$

This, in turn, shows how the adjustment factors $\alpha_i = e^{\beta_i}$ of the model then relate with claim severity. Indeed, the factors multiply together to make up an unbiased estimator of the average cost per case of the original claim population. The exponent c_i is the exponentially weighted average of the explanatory variable x_i .

Simulation Example

Two independent random variables x_2, x_3 are generated. The variable x_2 is a categorical variable satisfying:

$$\text{prob}(x_2 = 0) = \frac{1}{2} = \text{prob}(x_2 = 1)$$

The continuous variable x_3 is generated to conform to a normal density (slightly truncated):

$$x_3 \sim N(\log(40), \frac{3}{10}) \quad x_3 \in [0, 90]$$

The variable v is then generated also to conform to a normal density:

$$v \sim N(2 + \frac{x_2}{2}, \frac{1}{4})$$

Define

$$x_1 = w_1 \equiv 1, w_2 = x_2, w_3 = e^{x_3}, z = e^{v+x_3} \quad \text{and } y = \log(z)$$

Three weights are defined on W simulated observations $x = \langle x_1, x_2, x_3 \rangle$ according to the values taken on by $z = z_x$. The first weight ω is regarded as the "natural" sampling weight and is selected to be a uniform weight: $\omega_x \equiv 1$. The second weight γ is the exponentially adjusted weight and the third weight $\rho = \rho_x = z_x$ is the proportional cost weight.

Two generic models are considered. Model A is a linear cost model included for simple comparison purposes:

$$\text{A: } z = \eta_1 w_1 + \eta_2 w_2 + \eta_3 w_3$$

while Model B is a log cost model of the type considered in this paper:

$$\text{B: } y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = \log(z) = \beta_1 w_1 + \beta_2 w_2 + \beta_3 \log(w_3)$$

Observe that the data has been simulated to approximate Model B (with $\beta_1 = 2$, $\beta_2 = \frac{1}{2}$, and $\beta_3 = 1$). The first goal is to determine the sensitivity of these models to the choice of weight. The second goal is to determine how the choices of weights and models fair at estimating the mean dollar cost z, especially relative to the categorical explanatory variable w_2. The tables below show the parameter values from simulations of several population sizes. Standard errors are shown in parentheses.

Simulation of Regression Model A

Weight		$W = 100$	$W = 1000$	$W = 10000$	$W = 100000$	$W = 1000000$
ω	η_1	-48.928 (37.598)	-105.037 (13.774)	-98.682 (4.277)	-102.717 (1.349)	-103.064 (0.426)
	η_2	227.890 (21.547)	216.078 (7.796)	209.139 (2.398)	207.398 (0.758)	206.883 (0.240)
	η_3	8.073 (0.853)	9.999 (0.303)	9.984 (0.094)	10.073 (0.030)	10.087 (0.009)
	Adj R^2	0.678	0.654	0.655	0.655	0.656
	$\bar{z}$	396.27	421.09	423.74	422.06	422.20
γ	η_1	-55.648 (42.981)	-135.644 (15.651)	-118.705 (4.776)	-127.639 (1.514)	-127.441 (0.478)
	η_2	242.273 (23.919)	236.794 (8.864)	226.280 (2.673)	224.840 (0.850)	223.880 (0.269)
	η_3	8.291 (0.923)	10.795 (0.325)	10.572 (0.099)	10.769 (0.031)	10.777 (0.010)
	Adj R^2	0.656	0.649	0.651	0.652	0.653
	$\bar{z}$	440.32	470.75	471.12	469.42	469.80
ρ	η_1	-57.862 (48.639)	-167.369 (17.614)	-135.862 (5.314)	-151.328 (1.693)	-150.153 (0.534)
	η_2	259.884 (26.730)	258.331 (10.102)	243.405 (2.993)	242.514 (0.957)	241.057 (0.302)
	η_3	8.393 (0.990)	11.565 (0.345)	11.071 (0.104)	11.403 (0.033)	11.396 (0.010)
	Adj R^2	0.630	0.645	0.642	0.646	0.646
	$\bar{z}$	485.88	525.12	522.07	520.71	521.27

The negative values for the intercept term η_1 illustrate the problem of negative cost predictions by linear cost models.

Simulation of Regression Model B

Weight	Coefficient	$W = 100$	$W = 1000$	$W = 10000$	$W = 100000$	$W = 1000000$
ω	β_1	2.134 (0.283)	1.989 (0.097)	2.017 (0.031)	2.006 (0.010)	2.002 (0.003)
	β_2	0.579 (0.048)	0.517 (0.016)	0.503 (0.005)	0.500 (0.002)	0.500 (0.001)
	β_3	0.939 (0.077)	0.999 (0.026)	0.996 (0.008)	0.998 (0.003)	1.000 (0.001)
	Adj R^2	0.763	0.720	0.709	0.708	0.709
	$\bar{y}$	5.87	5.93	5.94	5.94	5.94
γ	β_1	2.270 (0.296)	2.031 (0.010)	2.055 (0.032)	2.030 (0.010)	2.024 (0.003)
	β_2	0.586 (0.048)	0.523 (0.016)	0.506 (0.005)	0.502 (0.002)	0.501 (0.001)
	β_3	0.908 (0.080)	0.995 (0.027)	0.994 (0.008)	1.000 (0.003)	1.002 (0.001)
	Adj R^2	0.747	0.716	0.707	0.709	0.709
	$\bar{y}$	5.98	6.04	6.05	6.05	6.05
ρ	β_1	2.417 (0.310)	2.081 (0.102)	2.111 (0.032)	2.069 (0.010)	2.063 (0.003)
	β_2	0.592 (0.050)	0.528 (0.017)	0.506 (0.005)	0.502 (0.002)	0.500 (0.001)
	β_3	0.875 (0.082)	0.990 (0.027)	0.987 (0.008)	0.998 (0.003)	1.000 (0.001)
	Adj R^2	0.724	0.709	0.701	0.704	0.704
	$\bar{y}$	6.09	6.15	6.16	6.15	6.15

The truncation of x_3 is observed to especially impact the intercept and, as one would expect, this impact is greater with a weight that is skewed toward higher (truncated) values of z. Even so, the results illustrate that when an OLS model is fairly well specified, like Model B, the choice of weight has little effect on the parameter values.

The following three tables show the results of using the regression equations to estimate a mean value for z, the preferred severity measure, both overall and according to the characteristic x_2. For model B there are two natural ways to do this. One is to first exponentiate individual predictions for y and then take their mean (using the natural weight ω). Another is to first average the individual predictions for y (using the same weight as that used in determining the regression equation) and then exponentiate the result. On an overall basis, the combination of Model A and weight ω guarantees complete accuracy. So too does Model B with weight γ and

the second method. The results are provided in the following tables. The choice of weight applies to the determination of the regression equations and to averages on the logged scale (average then exponentiate method). All actual and estimated mean costs are determined using the natural weight ω. Observe that for Model B and weight ω, the first method (exponentiating then averaging) yields a higher estimate than the second method--this is a consequence of the arithmetic weighted mean being greater than or equal to the geometric weighted mean:

exponentiate then average $\Downarrow$ average then exponentiate $\Downarrow$

$$\frac{1}{W}\sum \omega e^{\hat{y}} \geq [\prod e^{\omega \hat{y}}]^{\frac{1}{W}} = e^{\frac{1}{W}\sum \omega \hat{y}}$$

Model A Estimates

Wt	x_2	$W = 100$		$W = 1000$		$W = 10000$		$W = 100000$		$W = 1000000$	
		Act.	Est.	Act.	Est.	Act.	Est.	Act.	Est.	Act.	Est.
ω	0	276	276	312	312	318	318	319	319	319	319
	1	512	512	532	532	529	529	525	525	526	526
	All	397	397	421	421	424	424	422	422	422	422
γ	0	276	278	312	314	318	323	319	323	319	323
	1	512	532	532	556	529	551	525	547	526	547
	All	397	407	421	434	424	437	422	435	422	435
ρ	0	276	280	312	315	318	326	319	326	319	327
	1	512	551	532	578	529	571	525	567	526	568
	All	397	418	421	445	424	449	422	447	422	447

Model B Estimates-Exponentiate then Average Metho

Wt	x_2	$W = 100$		$W = 1000$		$W = 10000$		$W = 100000$		$W = 1000000$	
		Act.	Est.	Act.	Est.	Act.	Est.	Act.	Est.	Act.	Est.
ω	0	276	270	312	304	318	309	319	310	319	309
	1	512	498	532	514	529	512	525	509	526	509
	All	397	386	421	408	424	412	422	409	422	409
γ	0	276	276	312	312	318	318	319	319	319	319
	1	512	512	532	532	529	529	525	525	526	526
	All	397	396	421	421	424	424	422	422	422	422
ρ	0	276	283	312	321	318	328	319	329	319	329
	1	512	527	532	550	529	546	525	542	526	542
	All	397	407	421	435	424	437	422	435	422	436

Model B Estimates-Average then Exponentiate Method

Wt	x_2	$W = 100$		$W = 1000$		$W = 10000$		$W = 100000$		$W = 1000000$	
		Act.	Est.	Act.	Est.	Act.	Est.	Act.	Est.	Act.	Est.
ω	0	276	257	312	290	318	295	319	296	319	295
	1	512	481	532	492	529	490	525	487	526	487
	All	397	354	421	377	424	381	422	379	422	379
γ	0	276	276	312	312	318	318	319	319	319	319
	1	512	512	532	533	529	529	525	526	526	526
	All	397	396	421	421	424	424	422	422	422	422
ρ	0	276	294	312	335	318	343	319	344	319	344
	1	512	544	532	576	529	570	525	567	526	567
	All	397	441	421	470	424	471	422	469	422	470

As would be expected, use of alternative weight γ or ρ makes little sense for Model A. Indeed, both those weights skew toward higher cost cases and result in an estimate which consistently exceeds the actual cost. In general, the size of the simulation serves only to stabilize the figures. When using the log cost Model B to estimate mean costs, ω consistently understates and ρ consistently overstates. This conforms to earlier observations. It is interesting to observe that the exponentiate then average method generates closer to actual estimates using either ω or ρ. The key observation is that the exponential weight γ yields essentially the same estimates under either method, always being within 1 of the actual. These observations hold both overall and for the two categories determined by the characteristic x_2 .

This simple but fairly generic simulation suggests that whether the mean is taken before or after exponentiating, the exponential weight provides a considerable improvement in the ability of the logged cost regression equation to provide average severity estimates. It also points out that pragmatically, despite issues like conceptual interpretation and handling outliers, a logged cost model is not always preferable to a linear cost model.

Empirical Example

The DCI database is used to determine a logged model of incurred claim costs:

$$y = \Sigma \beta_i x_i \quad \text{where } y = \log(z) \text{ and } z = \text{cost}$$

As in the previous section, three weights are considered.. The first weight ω is the DCI sampling weight. The second weight γ is the exponentially adjusted weight and the third weight $\rho = \rho_x = z_x$ is the proportional cost weight. The following table summarizes the results. The model includes 109 explanatory variables, 40 of which are state dummy variables and another 20 which are industry group indicators not included in the table. Only two variables are continuous (log of pre-injury wage and log of claimant age at injury) the others are all {0=no,1=yes} characteristic variables. As above, the standard error of the coefficient estimate is shown in parentheses:

Explanatory Variable	ω		γ		ρ	
	Mean	Coefficient	Mean	Coefficient	Mean	Coefficient
Intercept	1.000	4.645 (0.044)	1.000	5.482 (0.044)	1.000	6.842 (0.044)
Fatal or PT case	0.004	0.960 (0.052)	0.012	1.365 (0.028)	0.031	1.238 (0.017)
Scheduled PPD case	0.079	0.770 (0.013)	0.133	0.515 (0.010)	0.177	0.271 (0.008)
Nonscheduled PPD case	0.030	0.742 (0.019)	0.053	0.497 (0.014)	0.075	0.252 (0.011)
Other PPD or lump sum award	0.032	0.867 (0.018)	0.053	0.612 (0.014)	0.071	0.342 (0.012)
2nd DCI report basis	0.217	0.582 (0.009)	0.270	0.675 (0.010)	0.307	0.713 (0.011)
3rd DCI report basis	0.055	1.258 (0.018)	0.108	1.381 (0.015)	0.174	1.361 (0.013)
4th DCI report basis	0.015	1.565 (0.037)	0.039	1.691 (0.026)	0.075	1.662 (0.020)
Open 1st Report	0.150	1.588 (0.010)	0.223	1.570 (0.009)	0.263	1.418 (0.009)
Open 2nd report	0.050	1.306 (0.016)	0.104	1.295 (0.012)	0.170	1.128 (0.010)
Open 3rd report	0.020	0.891 (0.028)	0.051	0.913 (0.018)	0.101	0.816 (0.014)
Open 4th report	0.007	0.872 (0.050)	0.023	0.827 (0.031)	0.051	0.677 (0.022)
Accident year 1995	0.269	0.063 (0.008)	0.270	0.057 (0.008)	0.274	0.023 (0.008)
Accident year 1996	0.243	0.185 (0.009)	0.245	0.170 (0.009)	0.247	0.136 (0.008)
Accident year 1997	0.166	0.141 (0.010)	0.176	0.097 (0.010)	0.174	0.085 (0.010)
Log of pre-injury wage	5.842	0.143 (0.004)	5.936	0.139 (0.004)	6.030	0.130 (0.004)
Log of claimant age at injury	3.565	0.414 (0.010)	3.607	0.376 (0.010)	3.642	0.279 (0.009)
Female claimant	0.298	-0.116 (0.008)	0.284	-0.152 (0.007)	0.259	-0.144 (0.007)
Hospital care	0.593	0.717 (0.007)	0.702	0.645 (0.007)	0.788	0.478 (0.007)
Vocational rehabilitation	0.025	0.880 (0.020)	0.059	0.703 (0.013)	0.109	0.506 (0.010)
Employer payroll from $1 to $100K	0.195	0.243 (0.012)	0.217	0.223 (0.012)	0.232	0.175 (0.011)
Employer payroll from $100K to $1M	0.270	0.201 (0.011)	0.277	0.199 (0.011)	0.286	0.187 (0.011)
Employer payroll from $1M to $10M	0.269	0.097 (0.011)	0.248	0.082 (0.011)	0.243	0.088 (0.011)
Employer payroll over $10M	0.163	0.082 (0.012)	0.158	0.048 (0.012)	0.150	0.024 (0.012)
Injury to chest	0.041	-0.012 (0.018)	0.036	-0.188 (0.018)	0.028	-0.446 (0.018)
Injury to head	0.038	-0.125 (0.018)	0.038	0.051 (0.017)	0.046	0.203 (0.015)

Injury to lower back	0.200	-0.028 (0.011)	0.202	-0.044 (0.011)	0.214	-0.116 (0.010)
Injury to lower extremity	0.219	-0.086 (0.011)	0.213	-0.171 (0.010)	0.203	-0.350 (0.009)
Injury to neck	0.018	0.055 (0.025)	0.021	-0.039 (0.022)	0.023	-0.194 (0.019)
Injury to upper back	0.020	-0.056 (0.024)	0.020	-0.079 (0.022)	0.019	-0.135 (0.021)
Injury to upper extremity	0.314	-0.065 (0.011)	0.306	-0.191 (0.010)	0.275	-0.388 (0.009)
Amputation	0.011	0.234 (0.033)	0.015	0.196 (0.027)	0.019	0.188 (0.023)
Burn	0.020	-0.285 (0.035)	0.016	-0.011 (0.034)	0.017	0.467 (0.030)
Carpal tunnel	0.021	0.189 (0.027)	0.027	0.022 (0.023)	0.027	-0.145 (0.021)
Concussion	0.102	-0.252 (0.014)	0.089	-0.270 (0.014)	0.078	-0.348 (0.013)
Fracture	0.122	0.021 (0.014)	0.145	0.015 (0.013)	0.167	-0.001 (0.011)
Infection	0.021	-0.059 (0.024)	0.020	-0.089 (0.023)	0.018	-0.192 (0.022)
Laceration	0.081	-0.282 (0.018)	0.064	-0.292 (0.018)	0.046	-0.364 (0.017)
Cumulative injury	0.021	-0.050 (0.025)	0.022	-0.084 (0.023)	0.021	-0.215 (0.022)
Strain	0.496	-0.142 (0.012)	0.482	-0.118 (0.011)	0.471	-0.154 (0.010)
Caused by burn	0.026	-0.315 (0.032)	0.021	-0.351 (0.030)	0.021	-0.278 (0.027)
Caused by being caught	0.048	-0.183 (0.019)	0.048	-0.164 (0.018)	0.048	-0.080 (0.017)
Caused by cumulative exposure	0.041	0.138 (0.021)	0.046	0.008 (0.019)	0.043	0.109 (0.018)
Caused by being cut	0.048	-0.190 (0.021)	0.038	-0.204 (0.021)	0.029	-0.214 (0.021)
Caused by a fall	0.215	0.009 (0.013)	0.236	-0.031 (0.012)	0.259	-0.047 (0.011)
Caused by an auto accident	0.039	0.040 (0.020)	0.052	0.038 (0.017)	0.072	0.085 (0.014)
Caused by straining	0.362	0.008 (0.013)	0.340	-0.071 (0.011)	0.312	-0.139 (0.011)
Caused by striking	0.049	-0.167 (0.018)	0.044	-0.194 (0.018)	0.038	-0.207 (0.017)
Caused by being struck	0.087	-0.154 (0.016)	0.081	-0.155 (0.015)	0.077	-0.109 (0.014)
N	xxx	112,841	xxx	112,841	xxx	112,841
Adjusted R^2	xxx	0.533	xxx	0.583	xxx	0.563
F-Value	xxx	1,195.41	xxx	1,463.67	xxx	1,348.12
Prob>F	xxx	0.0001	xxx	0.0001	xxx	0.0001
$\bar{y}$	8.25	xxx	9.379	xxx	10.43	xxx
$\bar{z}$	11,840	xxx	35,089	xxx	81,109	xxx

The following table summarizes the results of using the regression equations to estimate a mean value for z, the preferred severity measure, both overall and according to accident year. The same two methods are used as in the previous section: Method 1 exponentiates individual predictions for y and then takes their mean (using the natural weight ω) while Method 2 first averages the individual predictions for y (using the same weight as that used in determining the regression equation) and then exponentiates the result.

Accident Year	Actual Cost	ω		γ		ρ	
		Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
1994	11,274	7,286	3,427	11,398	11,203	17,642	61,325
1995	12,031	7,903	3,770	12,420	12,057	18,682	62,259
1996	12,000	7,883	3,945	12,276	11,998	18,787	54,826
1997	12,386	8,436	4,626	12,631	12,450	19,711	42,445
All	11,840	7,789	3,825	12,092	11,840	18,545	56,698

Again and as would be expected, ω consistently understates and ρ consistently overstates. It is again interesting to observe that Method 1 generates closer to actual estimates using either weight ω or ρ. The combination of Method 2 and the sampling weight estimates the average cost per case by the geometric mean, which grossly understates the severity measure. The combination of Method 2 and the proportional weight overstates the mark even worse. A key observation is again that the exponential weight γ yields markedly better results with essentially similar estimates under either method. The best results are obtained combining Method 2 with the exponential weight. Indeed, the estimated cost is observed to fall within 1% of the actual, both overall (where equality is prearranged) and by accident year.

Since the mean cost by accident year stays fairly close to the overall mean, the close to actual estimates with the exponential weight are not surprising. The following table is similar to the above but itemizes according to claim status. Open claims, on average, are more than double and closed cases only about half the overall mean:

Claim Status	Actual Cost	ω		γ		ρ	
		Method 1	Method 2	Method 1	Method 2	Method 1	Method 2
Open	30,423	20,971	15,026	32,117	31,538	45,976	81,006
Closed	6,367	3,906	2,556	6,194	6,136	10,519	22,491
All	11,840	7,789	3,825	12,092	11,840	18,545	56,698

While the estimates are not as close to the actual as for accident year, the exponential adjusted weight again does significantly better than the others. It is important to note that not only are those estimates closer to actual in dollar terms, but also in the relativity between open and closed. This indicates that the relativity varies with size of loss and a change in focus improves the logged cost regression severity estimates.

This result provides strong empirical evidence that using either Method 1 or 2, the exponential weight provides a considerable improvement in the ability of the logged cost regression equation to provide estimates for workers compensation case severity.

Conclusions

When it is desired to use a log-linear or log-log cost model to determine aggregate or mean costs, use of the exponential adjusted weight in the regression may provide a practical way to significantly improve the estimate The adjustment is most useful when there is the concern that the model parameters may vary with the size of loss. It exploits the fact that the regression model affords its best prediction at the center of gravity of the weighted data. The idea is to select a weight which focuses the log cost model on the most appropriate severity measure. Typically that measure will be expressed in the original dollar units and equal the average cost per case (weighted by probability of claim occurrence). The technique also improves the result of exponentiating the log cost model equation to produce individual or average claim cost estimates.

There is an evident asymmetry to the definition of exponential adjusted weight as defined here. This is due to the choice of working from smaller to larger costs. This "ascending" approach was arbitrary. One could equally well work in the other direction and this would result in a "descending" adjusted weight. It is easy to construct examples when the two differ. A weighted average of the two would provide a candidate for a more symmetric exponentially adjusted weight, but it is unclear whether that would add any value.

While this paper emphasizes the value of considering multiple alternative weights, the results suggest the idea of an "optimal weight" (or measure) relative to a properly specified problem. A practical example is presented of how the application of the log function may be paired with the use of an alternative weight. In the case in point, the distortion from a change in scale (the log transformation) is offset by a change in the center of gravity (the exponential weight). The methodology clearly applies to a transformation by any strictly monotonic function. Any reasonably smooth transformation could be handled by restriction to intervals over which the function is strictly monotonic (intervals on which it is constant do not require any weight shift), being sure to maintain the original weight distribution among those intervals. It may be useful to generalize this into a "change of (weight) variable rule" for transforming the dependent variable of a linear regression.

Appendix--SAS Routine to Calculate the Exponential Adjusted Weight

The following SAS steps begin with a data set named DCI which contains numeric variable Z and weight variable A. They end by including into DCI the exponential adjusted weight variable B, as defined in the paper (also, DCI is sorted by descending order of Z and any existing data sets named ONE, SONE or TWO are overwritten). The variable B is normalized to have the same total weight as the original weight variable A over the data set DCI. The two weights are defined to be proportional over subsets of observations with same value for the variable Z. The interested reader can readily modify this into a SAS macro.

```
PROC SORT DATA=DCI;BY DESCENDING Z;

DATA TWO;SET DCI;
KEEP Z A;

DATA ONE (KEEP=Z WGT) SONE (KEEP=SWGT SZWGT);
SET TWO END=EOF;BY DESCENDING Z;
IF FIRST.Z THEN WGT = 0;
WGT + A;
IF  LAST.Z THEN DO;
        OUTPUT ONE;
        SWGT + WGT;SZWGT + Z*WGT;
        IF EOF THEN OUTPUT SONE;
        END;

DATA ONE;SET ONE END=EOF;
KEEP Z WGT WGTB; RETAIN W SB SA SAZ;
IF _N_ = 1 THEN DO;
        SET SONE;
        SB = 0;SA = SWGT;SAZ = SZWGT;   W = SAZ/SA;
        END;
A = WGT;
IF NOT EOF THEN DO;
        W1 = (SA*W - A*Z)/(SA - A);
        B  = (1-SB)*((LOG(W) - LOG(W1))/(LOG(Z) - LOG(W1)));
        SB + B;   SA + (-A);   SAZ + (-A*Z);   W = W1;
        END;
             ELSE
        B  = 1-SB; WGTB  = SWGT*B;

DATA DCI;MERGE DCI(IN=IND) ONE (IN=INO);
BY DESCENDING Z;
IF IND & INO;
B = A*(WGTB/WGT);
```