Monitoring Cost Changes with Log-Linear Cost Models: Lessons from a Case Study

Dan Corro and Kyumin Shim

Monitoring Cost Changes with Log-Linear Cost Models: Lessons from a Case Study

By

Dan Corro* and Kyumin Shim**

* Director of Claims Research

** Research Economist

National Council on Compensation Insurance, Inc.
5 Marine View Plaza
Hoboken, New Jersey 07030-5722
Phone: 201-222-0500, extension 2130
E-Mail: Dan_Corro@NCCI.com

December, 2000

Abstract:

When studying Worker's Compensation (WC) claim cost experience, researchers often prefer models that relate claim characteristics and other cost drivers to the logarithm of the claim cost, rather than to the dollar cost itself. Linear models based directly on dollars, however, are better suited to decomposing the differences in costs observed over time or between claim populations. Reconciling the two methods within one analysis can be awkward. This led us to a new perspective: one that enables the two approaches to work together while preserving the most desirable features of each.

The paper presents a general method for analyzing cost differences. It also illustrates the method in the context from whence it came: monitoring the post-reform experience of WC claim costs.

Keywords: Workers' Compensation Insurance, reform, Oxacca decomposition, log-linear model, log-log model, exponential weight.

Introduction

Analysts are often asked to interpret the economic landscape and assess the influence of several exogenous or predetermined factors on one endogenous variable. An example is workers' compensation [WC] claim cost taken as the endogenous variable to be studied in reference to a list of exogenous claim characteristics and cost drivers. Models are associated with some sort of mathematical representation such as linear, nonlinear, logarithmic linear function form, etc. From the structural perspective, the coefficients (or derivatives, or elasticities) from the different models correspond to different interpretations. From the standpoint of statistical considerations, there are reasons to opt for one structural model over another if it enhances our ability to interpret the data. That model choice, however, may not prove convenient when those cost relationships are only

a part of a larger investigation. For example, it may be required to analyze how the average cost per case—not its logarithm-- has changed post-reform. This may demand some contortion to incorporate the model results into a picture suitable for decision-making. The need to fit a "round" cost model into a "square" hole within a summary report may lower the confidence level of those findings and raise the concern whether the methodology is internally consistent.

It is standard practice to use log-linear and log-log regression models in the analysis of WC claim costs. While useful for the investigation of proportional cost relationships, those transformed models are not well suited for predicting individual or even average dollar claim costs. Those models focus on the "geometric" mean cost while interest centers on the "arithmetic" average cost per case.

On the other hand, regression equations provide a powerful computational device for benchmarking select sets of claim costs and for analyzing dollar cost differences into components associated with cost drivers. This technique, based Oxacca style decompositions, exploits the fact that regression equations relate the "arithmetic" mean cost with average levels of the cost drivers.

This paper describes a method for changing the assigned weights of observations in the determination of the logged cost model. That "exponential weight" refinement is designed to improve the performance of the model after conversion back to a dollar scale. The derivation of a specific reweighting formula is motivated from the basic data fitting

geometry of OLS regression (see [1] where the technique is tested on a large database of WC lost time claims). The idea is just to shift the log-linear regression model from its "geometric" to an "arithmetic" perspective that makes it consistent with the decomposition formula.

The next three sections provide technical background material: (1) the use logged cost models, (2) Oxacca style difference equations and (3) the exponential weight. The next section outlines a general methodology for putting the three pieces together. This is illustrated in the final section that presents a case study. The case study deals with monitoring WC claim costs post reform and is the context from which this work evolved. An Appendix provides additional detail on regressions discussed in that case study.

The Use of Logged Cost Models

The use of log-linear and or log-log regression models is the preferred practice for the analysis of workers compensation insurance claim costs. For simplicity, we refer to regression equations in which the dependent variable is the logarithm of a dollar cost as "logged cost models". 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. 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 that 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 logged cost model. This is viewed as one of the prices to be paid for mitigating the influence of outlier claims. We pursue this from a simple geometric point of view rather than from the more challenging perspective of model specification error. We begin 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 typically occurs when costs are recalibrated via the log function. This suggests reweighting the data to offset that effect. Reweighting observations is a common practice in constructing regression models to temper the effect of outliers or more generally to deal with heteroscedasticity. In a subsequent section we introduce a reweighting scheme that shifts the focal point of a logged cost model so as to make it better suited to producing dollar cost estimates. We will show that from this weight's perspective, the advantages of the logged cost models can be essentially retained while generating figures more readily broken down into cost components.

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 logged cost models of the form:

$$Y = \ln(Z) = \sum \beta_i X_i + \varepsilon$$

where ε represents the error term. The X_i may be categorical or continuous and, if continuous, be expressed in their 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). For example, 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 has been argued, compromises the assumption of constant elasticity. There are, however, important considerations that challenge or at least mitigate that criticism. The point here is not to debate the issue but to simply point out that 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 logged cost model in this context is best seen 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 claim characteristics 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 researchers may cite a litany of more technical considerations, it is primarily 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 indicating whether an attorney represents the claimant. 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 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 Detailed Claim Information (DCI) database used in the case study discussed later, state specific sampling ratios are used. Also, DCI sampling rules require

that the claims be stratified 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 abuse 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. Let Γ denote a claim sample set. The set of weights $\{\omega_x \mid x \in \Gamma\}$ (which is really a function $\omega : \Gamma \to [0, \infty)$, but we ignore that nicety here) has 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 $W = \sum \omega_x$ 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 with the weighted mean coinciding with the expected claim cost:

$$E(Z) = \sum p_x z_x = \frac{\sum \omega_x z_x}{\sum \omega_x} = \frac{1}{W} \sum \omega_x z_x .$$

Oxacca Style Decompositions

Suppose the claim sample is divided into *n* mutually disjoint subsets:

$$\Gamma=\bigcup_{i=1}^{n}\Gamma_i \quad i\neq j\Rightarrow\Gamma_i\cap\Gamma_j=\phi$$

and consider a (weighted) ordinary least squares (OLS) linear model on the claim sample of the form:

$$Y=\sum_{i=1}^{n}\alpha_i\delta_i+\sum_j\beta_jX_j+\varepsilon \text{ where } \delta_i(x)=\begin{cases}0 & x\notin\Gamma_i\\ 1 & x\in\Gamma_i\end{cases}$$

We are interested in analyzing the differences of Y among these subsets akin to the Oxacca decomposition of mean differences from linear models. Let horizontal and vertical bars denote, respectively, taking a (weighted) mean and restriction to a subset. In this context, we may express the error term as:

$$\varepsilon=\varepsilon\cdot 1=\varepsilon\cdot(\sum_i\delta_i)=\sum_i\varepsilon_i \quad \text{where } \varepsilon_i=\varepsilon\cdot\delta_i$$

and a property of OLS regression implies that:

$$0=\bar{\varepsilon}=\overline{\varepsilon_i}=\overline{\varepsilon_{|\Gamma_i}},\quad 1\le i\le n$$

This leads us to Oxacca style decompositions of differences of means over the various subsets. Indeed, the differences can be itemized into "base" and "mix" components.

$$\overline{y_{|\Gamma_i}}-\overline{y_{|\Gamma_k}}=\underbrace{(\alpha_i-\alpha_k)}_{\text{base}}+\underbrace{\sum_j\beta_j\left(\overline{x_{j|\Gamma_i}}-\overline{x_{j|\Gamma_k}}\right)}_{\text{mix}}$$

It is important to keep in mind that these means are determined using the same weights as are used to determine the regression equation.

The base difference can be interpreted as "unexplained" in the sense that the cost model does not associate it with any claim characteristic other belonging to a particular subset.

Alternatively, it can be interpreted as the result of selecting a common "baseline claim", specified as a set of assumed values for the explanatory variables, and then using the cost model to generate two predicted costs for that same claim. The first assumes that the claim belongs to the first subset of the comparison and the second assumes it belongs to the second subset, all else equal. Subtracting the first predicted cost from the second determines the "difference in base cost" component.

It may be useful to further itemize the mix component, since its summands are related with the explanatory variables of the model. For example, we have referred to some of the explanatory variables as "claim characteristics" and to others as "cost drivers". The decomposition can effectively group together the set of marginal cost impacts associated with the covariates of the cost model.

The Exponential Weight

As was noted above, the translation to logarithms compresses costs and 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 that assigns more weight to higher cost claims when evaluating the regression model. For example, you could make the weight of an observation proportional to its dollar cost. It turns out, however, that such a weight overcompensates (c.f. [1]).

As before, let Z denote claim cost and begin with a set $\{\omega_z \mid z \in \Gamma\}$ of weighted costs from a claim sample of size N. We want to determine another set of N weights $\{\gamma_z \mid z \in \Gamma\}$ for

that same cost data that behaves better under taking logs. It turns out that there is an essentially unique way to do this—refer to [1] for details. The first step is to sort the data by size of cost $\Gamma = \{z_i \mid z_i \le z_{i+1}; 1 \le i \le N-1\}$. Simplify the notation by letting $\omega_i = \omega_{z_i}$ and $\gamma_i = \gamma_{z_i}$ denote the corresponding weights. There is an ordered set $\{\gamma_i \mid 1 \le i \le N\}$ called the *corresponding exponential weight* that is uniquely determined from the conditions:

$$\left(\prod_{i=1}^{k} z_i^{\gamma_i}\right)^{\sum_{i=1}^{k}\gamma_i} = \frac{\sum_{i=1}^{k}\omega_i z_i}{\sum_{i=1}^{k}\omega_i}; 1 \le k \le N \quad \text{and} \quad \sum_{i=1}^{N}\gamma_i = \sum_{i=1}^{N}\omega_i$$

This just means that the exponentially weighted geometric mean equals the weighted arithmetic mean determined using the original weight.

Putting the Pieces Together

This section presents the basic methodology in a simple but generic setting. All that is involved is putting the pieces together from the previous three sections. As above, we begin with a weight $\{\omega_z \mid z \in \Gamma\}$ and a decomposition

$$\Gamma = \bigcup_{i=1}^{n}\Gamma_i \quad i \ne j \Rightarrow \Gamma_i \cap \Gamma_j = \phi$$

Let γ_i be the exponential weight corresponding to the weight $\omega_{|\Gamma_i}$ on the sub-sample Γ_i. Combine the γ_i into a weight γ on Γ so that $\gamma_{|\Gamma_i} = \gamma_i$. Note that both weights ω and γ assign the same weight $W_i = \sum_{z \in \Gamma_i}\omega_z = \sum_{z \in \Gamma_i}\gamma_z$ to each sub-sample Γ_i

The weight γ provides the perspective that enables logged cost models to itemize differences among the sub-samples. To see this, we let $Y=\log(Z)$ as above. Also let a bar indicate the (weighted arithmetic) mean using the weight ω and a double bar the (weighted arithmetic) mean using the weight γ.

We are interested in how the cost Z changes over the Γ_i, as measured by the average cost per case that we denote by $\overline{z_i} = \overline{z_{|\Gamma_i}}$. Letting $r_{i,j} = \frac{\overline{z_i}}{\overline{z_j}}$ the idea is to decompose those relative differences in terms of explanatory variables.

So construct an OLS log-linear model using the weight γ:

$$\log(Z) = Y = \sum_{i=1}^{n} \alpha_i \delta_i + \sum_k \beta_k X_k + \varepsilon$$

We have arranged things so that

$$\log(\overline{z_i}) = \log\left(\frac{\sum_{z\in\Gamma_i} \omega_z z}{W_i}\right) = \log\left(\prod_{z\in\Gamma_i} z^{\frac{\gamma_z}{W_i}}\right) = \frac{\sum_{z\in\Gamma_i} \gamma_z \log(z)}{W_i} = \overline{\overline{y_{|\Gamma_i}}} = \alpha_i + \sum_k \beta_k \overline{\overline{x_{k|\Gamma_i}}}$$

and, as above, there is an Oxacca style decomposition:

$$r_{ij} = e^{\alpha_i - \alpha_j} \prod_k e^{\beta_k\left(\overline{x_{k|\Gamma_i}} - \overline{x_{k|\Gamma_j}}\right)}$$

$e^{\alpha_i - \alpha_j}$ = base cost compoent factor

$e^{\beta_k\left(\overline{x_{k|\Gamma_i}} - \overline{x_{k|\Gamma_j}}\right)}$ = factor associated with covariate X_k

This shows how to itemize the relative cost differences, expressed in dollar terms, using elasticities from a logged cost model.

The next section applies this when the claim sample is divided into four disjoint subsets.

Γ_1=TB, experience of a reform (Test) state pre-reform (Before)

Γ_2=CB, experience of a group of non-reform (Control) states pre-reform (Before)

Γ_3=TA, experience of a reform (Test) state post-reform (After)

Γ_4=CA, experience of a group of non-reform (Control) states post-reform (After).

As noted before, in that case study the covariates were grouped into two general categories: "claim characteristics" and "cost drivers". Those categories used to determine component factors associated with the explanatory variables of the log-linear cost model.

A Case Study: Monitoring Post Reform Claim Severity

Much of the previous discussion makes reference to this example. This final section illustrates the concepts discussed above. Along with revisiting the methodology, it discusses findings of some independent interest.

Background: NCCI post-reform monitoring (PRM) reports analyze losses in states those enacted major legislative reforms of their WC systems over the last decade. The reports attempt to gain an understanding of the effects of the reforms on the system outcomes, and evaluate the consistency of the outcomes with the reforms' objectives. With the availability of the necessary data, the post-reform monitoring reports compare the actual claim frequency and severity before the enactment of the laws with outcomes after. This section illustrates the analysis for a group of seven states (Arkansas, Connecticut, Florida, Georgia, Kansas, and Kentucky, Montana). These states enacted major legislative reforms from June 1, 1993 through July 1, 1994 and each was the focus of a post-reform study by NCCI during 1998. The paper *NCCI Post Reform Monitoring Reports* [2] provides background and presents findings for the same group of seven states within the context of post-reform cost analyses.

Data Source: The comparison of lost-time claim severity uses data from the NCCI Detailed Claim Information (DCI) database. The DCI is primarily used for research, and

contains detailed information on a stratified random sample of lost time claims. In addition to incurred and paid claim costs, the DCI includes many claim characteristics, such as the part of body injured, the nature of the injury and its cause. It also includes indicators for attorney involvement, vocational rehabilitation; claim milestones such as date of injury, date of first disability payment, return to work or claim closure; as well as claimant demographics like age, gender, and pre-injury wage. The post reform monitoring studies use multivariate cost models to control the mix of injuries, claim characteristics and claimant demographics and to evaluate average claim costs in the pre- and post-reform periods. Indices for medical costs and wages are used to hold purchasing power constant over the two time periods.

General Approach: The analysis compares average claim costs in the pre- and post-reform periods in the reform states with outcomes from a group of jurisdictions that did not enact major systemic reforms.[1] Workers compensation experience improved significantly during the time period considered here and that improvement was not confined only to states instituting statutory reforms.

While it is impossible to exactly isolate the effectiveness of reforms from the general turnaround in experience, it is important to evaluate reform within that broader context. A simple comparison of experience before and after reform cannot achieve this. To that end, the analysis incorporates the experience of a "control" group of states that did not enact major reforms. In comparing case severity of the "test" reform states to the non-

[1] Those states are : Alaska, Arizona, the District of Columbia, Idaho, Illinois, Indiana, Iowa, Louisiana, Maryland, Michigan, Mississippi, Missouri, South Carolina, Utah, Vermont, Virginia and Wisconsin.

reform states' experience, it is equally important to account for the fact that the respective mix of injuries can significantly influence the result.

Average claim costs are compared between the two time periods for the reform and control states. For the reform states, pre- and post-reform time periods were selected based on the effective date of the reform law (typically, the pre-reform period ran from 18 to 6 months before while the post-reform period ran from 6 to 18 months after). For the control group states, the pre-reform period used is June, 1992 to May, 1993 and the post-reform period is May, 1994 to April, 1995. Those periods were selected so that, on average, the injury dates would be aligned with the before and after periods in the reform states. Comparison of outcomes in the reform states with the non-reform states provides a reference to the industry trends, while still differentiating the reform and non-reform state experience.

Linear and Logged Cost Models: As discussed above, it is standard practice for researchers to model the logarithm of cost, $\log(Z)$, when building models of claim costs. It is however, comparatively rare to find a justification for this beyond an exercise in hand waving. Chart 1 below shows the actual incurred costs for the DCI claim sample, arranged by increasing cost. Each "actual" point represents one percentile of the cost. More precisely, the data is sequenced by increasing size of claim z and then collected into 100 subsets of approximately equal weight. Chart 1 also shows the corresponding mean of $\hat{z}$, the predicted cost using a linear cost model and a second fit using an analogous logged cost model.

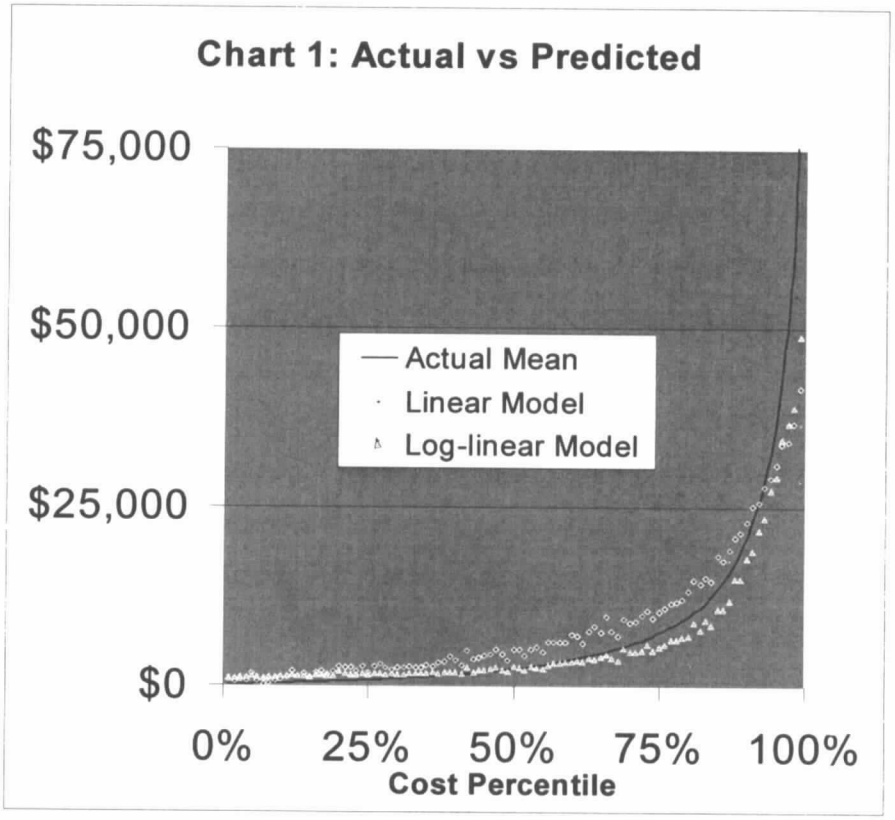


Chart 1: Actual vs Predicted
$75,000
$50,000
$25,000
$0
Actual Mean
Linear Model
Log-linear Model
0%
25%
50%
75%
100%
Cost Percentile

Predicted costs reflect regression toward the mean. Moreover, many of the explanatory values used in the cost models are $\{0,1\}$-indicator variables, which limits the range of predicted values. As a result, the fitted values show less variation than the actual costs. In particular, predicted costs understate the cost of the most expensive cases, a phenomenon that accounts for much of the error of the regressions. Chart 1 illustrates that while this is true for both linear and logged cost models, it is especially apparent for the linear model. Logged cost models typically exhibit a better fit. In this case, the adjusted R^2 is 0.983 for the logged cost model, more than double that of the linear model, at 0.427.

The graph of any (perhaps weighted) OLS linear model $z = f(x) + \varepsilon$ has a natural "center of gravity" at the point $\langle \bar{x}, \bar{z} \rangle = \langle \bar{x}, f(\bar{x}) \rangle$. When the same weight is used to construct a logged cost model $\log(z) = g_1(x) + \varepsilon$, however, the center of gravity of the regression, when transformed via exponentiation back to the original dollar scale, is moved to the point $\langle \bar{x}, \tilde{z} \rangle = \langle \bar{x}, \exp(g_1(\bar{x})) \rangle = \langle \bar{x}, \exp(\overline{\log(z)}) \rangle$ where $\tilde{z}$ is recognized as the (weighted) geometric mean of z. From the above remarks, we see that the sample weight can be "exponentially adjusted" in such a way that, when that new weight is used, the focal point of the logged cost model is shifted back to the (arithmetic) average cost per case. In this study, the exponential weight adjustment was applied to each of the four subsets $\{CB, CA, TB, TA\}$ identified above. Chart 2 compares the logged cost model fit using the sample and its corresponding exponential weight (refer to the Appendix for the logged cost model parameters using the exponentially adjusted weight).

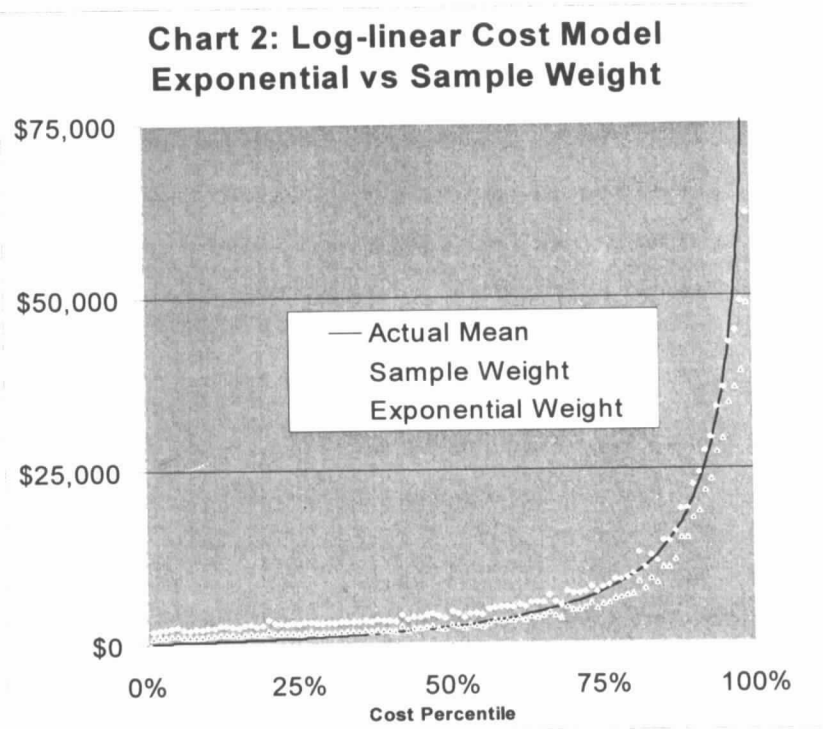


Chart 2: Log-linear Cost Model
Exponential vs Sample Weight
$75,000
$50,000
$25,000
$0
Actual Mean
Sample Weight
Exponential Weight
0%
25%
50%
75%
100%
Cost Percentile

Again, when weight is held constant, the effect of the logarithmic scale renders high cost z claims less influential in an OLS model for $\log(Z)$ than in an analogous model for Z. The exponential weight offsets that—whence its name—by assigning greater weight to the higher cost claims. This, in effect, shifts the center of gravity of the regression equation. Chart 2 illustrates this: while the sample weight log-linear fit is quite good from over 40-60th percentile range (the geometric mean of lost time costs is typically tracks with the median); the exponentially adjusted weight model fits best in the 70-90th percentile range (as is typical, the arithmetic mean of lost time costs—here about $10,000—is near the 80th percentile). The exponentially adjusted weight provides a better fit for high cost claims and optimizes the model fit near the value used to measure case severity. In this instance, the overall effect on the goodness of fit is small: use of the exponentially adjusted weight increases the adjusted R^2 slightly, to 0.988.

In light of the many $\{0,1\}$-indicator explanatory variables used in the cost models, it is worth recalling another advantage of logged cost models over simple linear models: most claim characteristics are more naturally associated with a proportional cost shift rather than a flat dollar loading. It should also be noted that continuous explanatory variables were converted to logarithmic scale in determining the logged cost models (log-log model form).

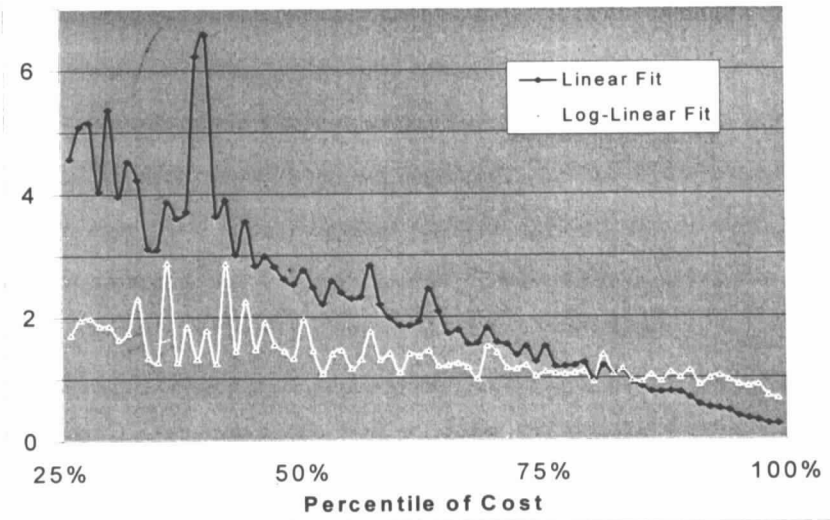
A more technical problem is that of heteroscedasticity. An important assumption of the classical OLS regression model $z = f(x) + \varepsilon$ is that the ε all have the same variance. As with much cross-sectional data, this is problematic in the case of WC case severity. Indeed, more expensive cases show greater cost variability and it is likely that this affects

the variability of the residuals. The presence of heteroscedasticity has important implications for the interpretation and application of the cost model, especially as regards predictions and their confidence intervals (its presence does not, however, invalidate the model coefficients used here to decompose cost differences). Although few would believe that lost time costs actually conform to any simple linear (or log-linear) functional form, in the classical OLS regression sense, this is relevant in light of the use the model to decompose cost differences. Indeed, the conceptual basis of the decomposition comes from interpreting the regression equation as the tangent hyperplane to the graph of the cost function at the center of gravity. The model coefficients regarded as partial derivatives that measure the slope at that point along the axis of the corresponding explanatory variable. The better the choice for the functional form of the cost model, therefore, the more credible the decomposition. By the same token, when using regression models to analyze case severity, it is advantageous to optimize the fit at a center of gravity which conforms to the severity measure being used—in this case the (sample weighted arithmetic) average cost per case.

Heteroscedasticity is also among the justifications cited for the use of the log transformation. The simplest approach to dealing with heteroscedasticity is to divide the observations into groups and examine the residuals for any pattern. Given the concern expressed above that higher cost cases are also the more variable, it is natural to again consider cost percentiles. Recall that in preparing Charts 1 and 2, claims were collected, according to size, into 100 groups of roughly equal weight. The idea here is to normalize the cost of each group to a common (weighted) mean of 1. The lowest quartile is excluded in order to avoid erratic results due, at least in part, to division by comparatively

small numbers. This generates 75 subsets of similar size and scale for which we can compare the model residuals. Chart 3 shows the standard deviation of the residuals for the linear and logged cost models, determined using the sample and exponentially adjusted weights, respectively (the pattern for the log-linear cost model derived using the original sample weight is quite similar to that using the exponential adjusted weight). Observe that, for both models, not only does the regression equation consistently under-predict the highest z values, it does so in such a way as to yield relatively little variation in the error, as compared with the size of z. While both models show a pattern of decline with increasing cost, that decline is less pronounced for the log-linear cost model. Indeed, while the log-linear variation measure remains mostly in the interval [1,2], the values from the linear model decline from 5 to nearly 0. From this simple picture, then, the log-linear cost model shows less evidence of heteroscedasticity.

Chart 3: Variation of Residual



To summarize, the case study illustrates the primary reason for using logged cost models is a much better fit to the data. Also, proportional cost effects are generally preferred to flat dollar loadings. Among the other motivations for using the log transformation is the need to counter heteroscedasticity and outliers by making higher cost cases less influential in the model. While the exponential weight adjustment runs somewhat counter to that by shifting weight to higher cost cases, it still improves the situation as regards heteroscedasticity and outliers and has the major advantage of optimizing the fit at the point measure of case severity.

Cost Decomposition: The previous two sections illustrate how convenient linear models are for decomposing dollar differences but that log-linear cost models generally provide a better fit to the data and have other conceptual advantages. This purpose of this section is again to put the pieces together. Applying the logarithm in conjunction with an "exponential" transformation of the sample weight, the mean values of the logged cost model invert back to the original (weighted) arithmetic mean. This enables a decomposition of the relative difference in case severity very similar to the Oxacca style dollar decomposition derived using linear cost models.

As above, the post-reform relative difference in mean cost per case among the non-reform states can therefore be expressed as:

$$\log(\overline{z}_{|CA}) - \log(\overline{z}_{|CB}) = \underbrace{(\alpha_{CA} - \alpha_{CB})}_{\text{base cost}} + \underbrace{\sum_j \beta_j \left(\overline{\overline{x_j}}_{|CA} - \overline{\overline{x_j}}_{|CB}\right)}_{\text{case mix}} + \underbrace{\sum_k \gamma_k \left(\overline{\overline{x_k}}_{|CA} - \overline{\overline{x_k}}_{|CB}\right)}_{\text{targeted cost drivers}}$$

This is the itemization of the relative difference in lost time case severity presented in the PRM studies. The results for the DCI claim data is shown in Tables 1a and 1b.

Table 1a: Components of Relative Difference: Post- vs Pre-Reform

Comparison Group	*Relative Difference #*	*Components*		
		Base Cost	*Claim Mix*	*Cost Drivers*
Control Group	-4.3%*	-13.3%	2.1%	6.9%
Test Group	-19.4%	-18.5%	2.6%	-3.6%

* Statistically different from 0 with 95% confidence, based on a 2-tailed T-Test.
# Relative difference of x Vs. y is determined as natural log(x/y), expressed as a percentage.
SOURCE: NCCI DCI, claims evaluated 18-months after report of injury.

Observe that for the reform states test group the cost drivers contributed to the decline in case severity, while those factors worked to increase costs in the non-reform states.

Table 1b: Components of Relative Difference: Test vs Control

Time Period	*Relative Difference #*	*Components*		
		Base Cost	*Claim Mix*	*Cost Drivers*
Pre-Reform	30.8%*	14.8%	-0.1%	16.1%
Post-Reform	15.7%*	9.6%	0.5%	5.6%

* Statistically different from 0 with 95% confidence, based on a 2-tailed T-Test.
# Relative difference of x Vs. y is determined as natural log(x/y), expressed as a percentage.
SOURCE: NCCI DCI, claims evaluated 18-months after report of injury.

The claim mix component is small in comparison with the other two components. This decomposition indicates that pre-reform cost drivers contributed a larger share to the higher severity of the reform states. The higher cost differential was cut in half post-reform and under this decomposition, targeted cost drivers account for a smaller share of that smaller difference.

Conclusions: A number of states enacted major reforms of their workers compensation systems in the last decade to control rapidly increasing claim frequency and costs. The most common tools to address these problems were the introduction of managed care provisions, the imposition of stricter compensability standards and fewer incentives for attorney involvement. NCCI post-reform monitoring reports analyze claim frequency and severity in these states before and after the enactment of reforms, comparing the outcomes to trends in a group of non-reform states. This paper describes the method used to analyze the severity of lost time cases using DCI claim data.

Factors other than the reforms, including the influence of economic cycles and secular trends, may have affected the outcomes. These factors may have countered the effects of the reforms where the observed improvements were modest. In addition, the analysis did not evaluate the impact of each reform provision on lost time case severity. It is likely that some reform measures may have greater impact than the others. For these reasons, a comparison of outcomes, such as a simple T-test of means, between the two periods with a reference to the countrywide trend provides only a limited understanding of the effects of the reforms on the system costs. As described here, multivariate cost models address this by decomposing the difference into components. A customized logged cost model is described and shown to possess some important technical features. That is the method used to prepare the PRM studies. The DCI results presented to illustrate the methodology indicate that cost drivers targeted by reform indeed play a different role in the reform states than in the non-reform control group of states. Still, those findings confirm the

view that factors other than those associated with claim characteristics captured in the DCI—like economic cycles and secular trends--may significantly influence costs.

From the reform versus non-reform state perspective, simple cost comparisons indicate that the reform states maintain a significantly higher case severity. That cost differential, however, was halved post-reform and the multivariate analysis assigns much of that relative improvement in claim severity to cost drivers targeted by reform

References

[1] Corro, Dan, 1999, A Practical Suggestion for Log-Linear Workers Compensation Cost Models. *Casualty Actuarial Society Forum. Spring 1999:* pp. 363-393.

[2] Corro, Dan and Helvacian, N. Mike, 1999, NCCI Post Reform Monitoring Reports. *NCCI 1999 Workers Compensation Mid-Year Issues Report:* pp. 11-15.

APPENDIX: Regressions Discussed in the Case Study

Dependent Variable: INCURRED COST

Table 1. Analysis of Variance

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	49	1.5142475E14	3.0903009E12	581.754	0.0001
Error	38145	2.0262767E14	5312037561.2		
U Total	38194	3.5405242E14			

Root MSE	72883.72631	R-square	0.4277
Dep Mean	8557.63163	Adj R-sq	0.4270
C.V.	851.68104		

Table 2. Parameter Estimates

Variable Description	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
TEST BEFORE SUBGROUP	1	-476.835672	729.14915362	-0.654	0.5131
TEST AFTER SUBGROUP	1	-2294.367470	747.57602381	-3.069	0.0021
CONTROL BEFORE SUBGROUP	1	-2080.279861	689.14790226	-3.019	0.0025
CONTROL AFTER SUBGROUP	1	-3001.570552	697.96546494	-4.300	0.0001
EMPLOYER PAYROLL SIZE $0	1	1469.861287	431.95451854	3.403	0.0007
EMPLOYER PAYROLL SIZE $1-$100K	1	557.942641	311.42257706	1.792	0.0732
EMPLOYER PAYROLL SIZE $100K-$1M	1	-6.767765	270.72389656	-0.025	0.9801
EMPLOYER PAYROLL SIZE $1M-$10M	1	240.760277	262.07125972	0.919	0.3583
CLASS IN SCHEDULE GROUP 05[2]	1	654.826799	570.71037800	1.147	0.2512
CLASS IN SCHEDULE GROUP 07	1	909.315255	1023.9962096	0.888	0.3745
CLASS IN SCHEDULE GROUP 10	1	105.300221	674.23379474	0.156	0.8759
CLASS IN SCHEDULE GROUP 12	1	-462.269804	674.63152253	-0.685	0.4932
CLASS IN SCHEDULE GROUP 14	1	93.132284	713.15373065	0.131	0.8961
CLASS IN SCHEDULE GROUP 17	1	185.409428	426.85143513	0.434	0.6640
CLASS IN SCHEDULE GROUP 18	1	-337.910290	492.17394439	-0.687	0.4924
CLASS IN SCHEDULE GROUP 20	1	-569.888682	957.46087475	-0.595	0.5517
CLASS IN SCHEDULE GROUP 21	1	572.977130	1546.8026243	0.370	0.7111
CLASS IN SCHEDULE GROUP 24	1	-102.002622	1141.1471662	-0.089	0.9288
CLASS IN SCHEDULE GROUP 25	1	1433.375913	1085.8342328	1.320	0.1868
CLASS IN SCHEDULE GROUP 26	1	801.569830	654.71935752	1.224	0.2208
CLASS IN SCHEDULE GROUP 27	1	1290.795104	379.54043241	3.401	0.0007
CLASS IN SCHEDULE GROUP 33	1	623.471859	1173.2744182	0.531	0.5951
CLASS IN SCHEDULE GROUP 34	1	-446.411085	299.33417378	-1.491	0.1359
CLASS IN SCHEDULE GROUP 35	1	432.100881	340.52812259	1.269	0.2045
CLASS IN SCHEDULE GROUP 36	1	-654.397301	343.07361011	-1.907	0.0565
TRAUMATIC INJURY	1	1034.054233	471.84163281	2.192	0.0284
PRE-INJURED WEEKLY WAGE	1	8.369938	0.52696165	15.883	0.0001
INJURY AGE	1	52.989464	7.92956518	6.683	0.0001
MALE CLAIMANT	1	1545.291364	223.90733173	6.901	0.0001
INJURED PART OF BODY = INTERNAL ORGANS	1	-4979.403718	564.72830329	-8.817	0.0001
INJURED PART OF BODY = HEAD	1	-373.060091	574.04135302	-0.650	0.5158
INJURED PART OF BODY = NECK	1	3235.641607	724.93327242	4.463	0.0001
INJURED PART OF BODY = LOWER BACK	1	-795.568308	335.14611281	-2.374	0.0176
INJURED PART OF BODY = UPPER BACK	1	-1479.748378	601.39009721	-2.461	0.0139
INJURED PART OF BODY = LOWER EXTREMITY	1	-2697.935157	337.36867599	-7.997	0.0001
INJURED PART OF BODY = UPPER EXTREMITY	1	-3309.790946	319.73928384	-10.352	0.0001
FATAL CLAIM	1	110398	3559.8311671	31.012	0.0001
STATUS OF CLAIM IS OPEN	1	24269	305.71886613	79.380	0.0001
WEEKLY BENEFIT	1	0.176502	0.06403396	2.756	0.0058
HOSPITALIZATION INDICATOR	1	3362.047743	199.67950657	16.837	0.0001
SURGERY INDICATOR	1	7044.530354	266.94389548	26.399	0.0001
VOCATIONAL REHABILITATION BENEFITS	1	25215	760.93884532	33.136	0.0001
CLAIMANT REPRESENTED BY AN ATTORNEY	1	3530.486859	305.37792234	11.561	0.0001
RETURN TO WORK INDICATOR	1	-3675.050427	204.25644599	-17.992	0.0001
PERMANENT TOTAL AWARD	1	75476	2277.5335650	33.139	0.0001
SCHEDULED PERMANENT PARTIAL AWARD	1	4859.151359	380.87552656	12.758	0.0001
NON-SCHEDULED PERMANENT PARTIAL AWARD	1	7546.803456	499.52747703	15.108	0.0001
DISFIGUREMENT AWARD INDICATOR	1	4956.894055	870.70708865	5.693	0.0001
LUMP SUM PAYMENT INDICATOR	1	10976	661.99166040	16.580	0.0001

[2] The classifications have been arranged into general industry divisions, designated "Schedules," and further subdivided into smaller "Groups" of classifications having similar or related characteristics. Source: Classification Codes & Statistical Codes for Workers' Compensation & Employers Liability Insurance, National Council on Compensation Insurance, Inc., 1997 Edition.

Dependent Variable: LOG OF INCURRED COST

Table 3. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	49	51997071.56	1061164.7257	65684.008	0.0001
Error	38145	616255.46055	16.15560		
U Total	38194	52613327.021			

Root MSE	4.01940	R-square	0.9883
Dep Mean	9.04757	Adj R-sq	0.9883
C.V.	44.42522		

Table 4. Parameter Estimates

Variable Description	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
TEST BEFORE SUBGROUP	1	4.930303	0.07994791	61.669	0.0001
TEST AFTER SUBGROUP	1	4.745682	0.08107615	58.534	0.0001
CONTROL BEFORE SUBGROUP	1	4.782400	0.07902394	60.518	0.0001
CONTROL AFTER SUBGROUP	1	4.649294	0.07995381	58.150	0.0001
EMPLOYER PAYROLL SIZE $0	1	0.146117	0.02352213	6.212	0.0001
EMPLOYER PAYROLL SIZE $1-$100K	1	0.071917	0.01710374	4.205	0.0001
EMPLOYER PAYROLL SIZE $100K-$1M	1	-0.027990	0.01514450	-1.848	0.0646
EMPLOYER PAYROLL SIZE $1M-$10M	1	-0.028072	0.01459912	-1.923	0.0545
CLASS IN SCHEDULE GROUP 05	1	0.111370	0.03093370	3.600	0.0003
CLASS IN SCHEDULE GROUP 07	1	-0.005698	0.06112679	-0.093	0.9257
CLASS IN SCHEDULE GROUP 10	1	0.010150	0.03737508	0.272	0.7860
CLASS IN SCHEDULE GROUP 12	1	0.031146	0.03685414	0.845	0.3980
CLASS IN SCHEDULE GROUP 14	1	0.078072	0.04032587	1.936	0.0529
CLASS IN SCHEDULE GROUP 17	1	0.074217	0.02269132	3.271	0.0011
CLASS IN SCHEDULE GROUP 18	1	0.048409	0.02663034	1.818	0.0691
CLASS IN SCHEDULE GROUP 20	1	0.091989	0.05046347	1.823	0.0685
CLASS IN SCHEDULE GROUP 21	1	0.061793	0.08601793	0.718	0.4725
CLASS IN SCHEDULE GROUP 24	1	0.020143	0.06040751	0.333	0.7388
CLASS IN SCHEDULE GROUP 25	1	0.331156	0.05723375	5.786	0.0001
CLASS IN SCHEDULE GROUP 26	1	0.123215	0.03439618	3.582	0.0003
CLASS IN SCHEDULE GROUP 27	1	0.094100	0.01950895	4.823	0.0001
CLASS IN SCHEDULE GROUP 33	1	0.067074	0.05952280	1.127	0.2598
CLASS IN SCHEDULE GROUP 34	1	-0.033787	0.01687428	-2.002	0.0453
CLASS IN SCHEDULE GROUP 35	1	0.074462	0.01862132	3.999	0.0001
CLASS IN SCHEDULE GROUP 36	1	-0.037061	0.02049899	-1.808	0.0706
TRAUMATIC INJURY	1	-0.107969	0.02368974	-4.529	0.0001
PRE-INJURED WEEKLY WAGE	1	0.129590	0.00814714	15.906	0.0001
INJURY AGE	1	0.309847	0.01681119	18.431	0.0001
MALE CLAIMANT	1	0.105804	0.01270655	8.327	0.0001
INJURED PART OF BODY = INTERNAL ORGANS	1	-0.404079	0.03242312	-12.463	0.0001
INJURED PART OF BODY = HEAD	1	-0.044888	0.03109403	-1.444	0.1489
INJURED PART OF BODY = NECK	1	0.159659	0.03573244	4.468	0.0001
INJURED PART OF BODY = LOWER BACK	1	-0.058334	0.01772142	-3.292	0.0010
INJURED PART OF BODY = UPPER BACK	1	-0.108579	0.03453432	-3.144	0.0017
INJURED PART OF BODY = LOWER EXTREMITY	1	-0.250969	0.01801989	-13.927	0.0001
INJURED PART OF BODY = UPER EXTREMITY	1	-0.261498	0.01691327	-15.461	0.0001
FATAL CLAIM	1	2.050513	0.08586028	23.882	0.0001
STATUS OF CLAIM IS OPEN	1	1.487664	0.01239017	120.068	0.0001
WEEKLY BENEFIT	1	0.260116	0.00969333	26.834	0.0001
HOSPITALIZATION INDICATOR	1	0.708042	0.01237411	57.220	0.0001
SURGERY INDICATOR	1	0.586466	0.01220345	48.057	0.0001
VOCATIONAL REHABILITATION BENEFITS	1	0.808863	0.02547978	31.745	0.0001
CLAIMANT REPRESENTED BY AN ATTORNEY	1	0.373587	0.01376672	27.137	0.0001
RETURN TO WORK INDICATOR	1	-0.397740	0.01090690	-36.467	0.0001
PERMANENT TOTAL AWARD	1	1.415209	0.06972410	20.297	0.0001
SCHEDULED PERMANENT PARTIAL AWARD	1	0.644809	0.01624483	39.693	0.0001
NON-SCHEDULED PERMANENT PARTIAL AWARD	1	0.783743	0.02121763	36.938	0.0001
DISFIGUREMENT AWARD INDICATOR	1	0.597890	0.03679728	16.248	0.0001
LUMP SUM PAYMENT INDICATOR	1	1.138939	0.02717348	41.914	0.0001