Econometric Modeling of Insurance Frequency Trends: Which Model Should We Choose?

Amin Ussif, Ph.D.

Econometric Modeling of Insurance Frequency Trends: Which Model Should We Choose?

Amin Ussif[1], Ph.D.
223 6th Street
Fairview, New Jersey 07022
aussif@hotmail.com

Abstract

In policymaking and insurance rate setting process, understanding and managing claim frequency are crucial issues. Owing to the importance attached to the dynamics of claims frequency in insurance ratemaking and in implementing workplace safety measures, we intend to walk through the basic steps in the econometric modeling and forecasting of claims frequency. Data from the California Workers Compensation Institute (CWCI) are used in this study. Three competing models are developed with the goal of selecting a superior one amongst the three. All three specifications confirm the prior finding of the CWCI that economic activity is a significant determinant of workers compensation frequency. The conclusion is that the nonlinear models, (constant elasticity and the exponential or growth models) perform better than the linear model. Also, applying the likelihood ratio test and the F-test to the Actuarial models against the Econometric models, it is shown that considerable statistical gains can be achieved by using economic variables in estimating trends.

[1] I am grateful to Dan Corro and Greg Engl of NCCI, and Dr. Rashid Sumaila, Dr. Steve Morey and Betsy Fadali for their help.

Introduction

The goal in this paper is to investigate the effects of different functional form specification in modeling workers compensation claim frequency. While the economic theory of claims frequency is unambiguous, it is still not very clear how the frequency of claims filing is functionally related to the explanatory variables or covariates. One of the basic assumptions of linear regression is that the model is correctly specified thus making the choice of functional forms an important step in econometric modeling of claims frequency.

Understanding basic functional relationships is in fact very critical in the application of econometric modeling in practice. Forecasting is another very crucial aspect of insurance business, economics and finance. The health of the insurance industry depends on the accuracy of the forecasts. In setting premium rates, losses are forecast in advance and then rates are determined to cover claims when they occur.

This paper has a dual focus. First, it investigates the relationship between frequency and two key economic variables; employment and the unemployment rate. We use three different functional form specifications to test the various hypotheses about the impact of economic activity. The first functional form is the linear model, the second is the multiplicative model (a generalization of the Cobb-Douglas production function in economics) and the third is the exponential or growth model called the semi-log model in the econometrics literature.[2] The performance of these models is studied in order to objectively select the superior functional form based on statistics. The selection is based solely on information from the data and very little judgment is applied in order to maintain objectivity. Although the selection will be based on insample information, the real test of the quality of the models will be determined when we compare the predictive power of the models against experience outside the observation window. To our knowledge, this kind of study is the first ever that is conducted using the quarterly data from CWCI. Finally, we compare trend indications prior to credibility using the commonly employed models by actuaries, i.e. linear and exponential trend models. These actuarial trend models are special cases of the econometric counterparts developed in this paper.

---

[2] The second specification is the model developed by Kahley (2000) for the forecasting of frequency of claims for the California Workers Compensation Institute (CWCI).

The first section discusses the theoretical basis of the models. The economic theory of claims frequency filing is also presented and discussed. It then builds three mathematical models allowing for relevant nonlinearities. Several practical issues are also addressed. The final section discusses the empirical results and concludes the paper.

The Economic theory of Claims Frequency[3]

In general, it will be assumed that frequency is a function of employment, unemployment rate, and a trend. The basic hypothesis is that economic activity is a determinant of workers compensation claims frequency, i.e. increases in economic activity lead to increases in frequency holding other factors constant. It is also hypothesized that there are other ways that the economic environment affects the claim filing activity apart from the effects on the incidence of injuries. For example, the availability of jobs and the health of the labor market as reflected by changes in unemployment rate, plant closures, layoffs, etc., are potentially important causes of the incidence of claim filing (Kahley 2000). Another plausible postulate is that certain variables such as technology, safety initiatives from employers, etc impact the accident rates (see Ussif, 2002). For instance, technical progress and increases in workplace safety will a priori reduce incidence rates. We now specify various functional forms based on the above hypothesis.

Functional Form Specifications

Three alternative econometric models are considered in this section. All models have the same number of explanatory variables but differ only in their functional form specification. The first specification is linear in the variables while the others are nonlinear. The econometric models are given by the following equations

$$Y_t = \beta_1 + \beta_2 Emp_t + \beta_3\, Unemp_t + \beta_4 Time + \varepsilon_t \qquad (1)$$

$$Y_t = \beta_1 Emp_t^{\beta_2} Unemp_t^{\beta_3} \exp(\beta_4 Time + \varepsilon_t) \qquad (2)$$

$$Y_t = \exp(\beta_1 + \beta_2 Emp_t + \beta_3 Unemp_t + \beta_4 Time + \varepsilon_t) \qquad (3)$$

---

[3] Frequency is defined as number of claims per earned premium. Please see Kahley for more about the data.

where exp(.) is the exponent function, *Emp* is the employment, *Unemp* is the unemployment rate and *Time* is the time trend dummy. Equation (2) is what is usually called the logarithmic-linear model while equation (3) is the semi-logarithmic model. The log-linear model has the advantage that the coefficients are the partial elasticities with respect to the independent variables. They simply tell us that a one percentage change in the independent variable will result in a certain percent change in the dependent variable. It is also clear that this model produces the average frequency growth rate frequency as the partial derivative of the dependent variable with respect to time. Equation three is the so-called exponential or growth model and its partial derivative with respect to time also gives the average frequency growth rate. These models extend the actuarial trend model to include economic covariates the unemployment rate and employment. By including these explanatory variables, the chances of capturing turning points may be greatly enhanced. To reiterate, these models may have some additional forecasting ability because of the information they used from the additional explanatory variables.

An important distinguishing feature of the models is that, the coefficient of the time trend variable in the linear model yields the absolute change in frequency per unit of time while the log-linear and the exponential give the percent change per unit of time. Hence, the nonlinear models have the additional advantage that they produce the equivalent of the actuarial trend estimates automatically.

In the application of econometric modeling to test refutable hypothesis and in forecasting, an important question is what makes a model "good"? To answer this question, we state a few criteria often used to help judge the "quality" of a model.

- Parsimony: - A mathematical model is a simplification of reality. It is not meant to capture all minor and random events but rather the essence of the phenomenon. All things being equal simpler models are preferred to unnecessarily large models. Simplicity in this context refers to the number of regressors and functional form.
- Theoretical consistency: - The coefficients in a model should have the right signs. A model may not be good if one or more of the estimated coefficients have the wrong signs. This has an important implication when using the model for purposes of forecasting.
- Goodness of fit: - A high adjusted R-square is good but this should not be overemphasized. Note that a model may not be good despite a high R-square if the estimated coefficients do not all have the

right signs. The main goal should not be to maximize the R-square but a good model with a high R-square is always welcome.

- Predictive power: - A good test of the validity of a model is comparison of its forecast with experience, its postsample predictive power. This also underscores the fact that a high fit does not necessarily mean good forecasting ability.

In practice, it is important to consider some of these criteria as guide towards consistent and reasonable forecasts.

Interpretation of results

Several statistics are used to explain in a relatively simple terms the necessary steps in using econometric analysis to help in policymaking and to enhance the understanding of claim frequency variable in insurance ratemaking.

It is obvious from the results (see Tables below) that the models have all performed reasonably well given the simplicity of their functional forms. The economic variables all have positive slope coefficients and are statistically significant at the 5% level of significance. That is a general increase in payroll which is normally a function of economic activity will lead to an increase in expected frequency. The positive sign on the unemployment rate is as expected since it reflects the conjecture that workers tend to file more claims during hard times in the labor market. Kahley (2000) provides some reasons to support this in California. This is a question of moral hazard and can be significant where the unemployment benefits are relatively low compared to the workers compensation benefits. In general, the trend variable has a negative sign and it is statistically significant which means that there was a long-run downward tendency in frequency in California. This may be attributable to factors such as safety measures, technical progress, etc. The meaning and practical application of the coefficients on the trend variable will be discussed in more detail later in this paper.

Table 1: Results of Indemnity Claims Frequency. Note that employment is in 100 000 workers.

Models Coefficients for Indemnity			
	Linear	*Multiplicative*	*Exponential*
Intercept	-135.82(-2.67)	-66.7100(-4.32)	-0.2750(-.27)
Employment	1.6(3.54)	4.2510(4.54)	0.028(3.95)
Unemp Rate	3.16(4.39)	0.5398(6.07)	0.0726(5.27)
Trend	-1.06(-7.07)	-0.0259(-9.56)	-0.0246(-8.41)
DW	1.9567	1.9451	1.9385
Adj. R^2	0.66	0.7928	0.7491
AIC	193.04	-106.85	-104.61
SBC	201.36	-98.53	-96.30

Table 2: Results of Medical Claims Frequency

Models Coefficients for Medical			
	Linear	*Multiplicative*	*Exponential*
Intercept	-127.94(-3.57)	-46.35(-5.35)	1.4235(-.27)
Employment[4]	1.58(6.24)	3.0925(5.88)	0.22(3.95)
Unemp Rate	0.5943(1.40)	0.0892(6.07)	0.0129(5.27)
Trend	-1.3918(-14.89)	-0.020(-14.18)	-0.020(-14.40)
DW	1.9907	1.9835	1.9906
Adj. R2	0.9381	0.9316	0.9323
AIC	202.79	-113.91	-113.91
SBC	220.11	-105.60	-105.60

The models have a high within sample predictive or explanatory power. The coefficient[5] of determination is used to judge the explanatory power of the regressors. For the time period considered, the economic factors together with the time trend explained about 67-95% of the variation in

[4] Employment in 100 000 workers.

[5] The implicit R-squares are calculated for the nonlinear models to make them comparable since the dependent variables are not the same.

frequency. It is clear that the linear model has the lowest explanatory power while the log-linear and semi-log are almost indistinguishable. Also, the R-squares are generally higher for the medical frequency compared to the indemnity frequency. After correcting for first order serial correlation[6], the DW statistic improved significantly in all the models. They are all close to 2.0 which is an improvement from the barely 1.0 before correction. In general, it appears that serial correlation is a menace in claims or frequency data.

The interpretation of the regression coefficients is also a very important part of econometric modeling. From the table of results (Table 1 and 2), for the linear specification, a one unit change in employment (unit is 1000 employees) holding other factors constant will result in 0.02 unit change in both indemnity and medical frequency. Also, for unemployment rate, a unit (%) change will result in a 3.16 unit change in frequency. The trend variable is negative indicating a small but persistent downward development in frequency. It may mean that over time, claims tend to decline due to improvements in factors such as technology and the manufacturing/service mix of the labor market.

Since the specification of model 2 automatically yields percent changes, the interpretation of the constant elasticity model is that a one percentage change in employment will result in respectively 4% and 3% changes in indemnity and medical frequency. The trends are discussed later in the paper. The interpretation of the exponential model requires some special attention. Note that, a unit change in employment will result in 0.0002 % change in frequency for both the indemnity and medical frequency. It is easy to see that, the coefficient measures the relative change in frequency for a unit absolute change in the independent variables.

Model Selection Techniques

In practice, e.g., in actuarial trending procedures, one is often saddled with the question of which model is preferable to some other model(s). According to the ASP, the actuary should 'be familiar with and consider various methods in statistics and numerical analysis for measuring trends". This also entails steps for evaluating the tentatively selected model and possibly revising the model. This in fact means that the actuary is not opposed to new and improved methods of model selection.

---

[6] Serial correlation is when errors in one time period are correlated directly with errors in ensuing period.

There are many statistics that may help make an objective and consistent selection among competing models. The adjusted R-square has often been used to select models for forecasting purposes. This statistic has sometimes been criticized for not adequately imposing adequate penalty for the degrees of freedom. Thus some modern criteria such as the Akaike Information (AIC) and Schwartz Bayesian criteria (SBC) have been proposed. In employing these criteria for judging model's performance, the smaller the value of the statistic the better. We shall use different analytical model selection criteria in deciding which model is best for forecasting. From the tables (1 and 2) of results, it is again clear that the two nonlinear models are the winners, i.e. both the log-linear and the exponential models have smaller AIC and SBC than the linear model. The log-linear model has the smallest AIC and SBC. Note that, it is generally accepted that when AIC and SBC conflict, one should choose the model with lowest SBC.

Forecasting Frequency

At several levels of insurance business, decisions have to be made. To guide decision makers forecasts are often produced. For example, forecasts of expected claims (pure premium) are required in making rates. Under the credibility approach, the premium estimate for a loss if full credibility is applied is the average loss from the experience. Forecasts of trends have always been used as inputs in ratemaking process.

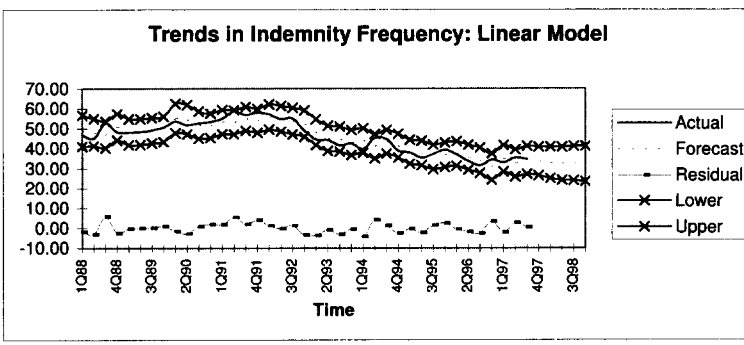
Understanding the steps in obtaining reasonable models is vital to improving the quality of the predictions and their application in real world. In light of these compelling reasons, we attempt to briefly explain the procedure using data from the California Workers Compensation Institute. To put our models to test, two types of forecasts are performed that is ex-post and ex-ante predictions. In the ex-post forecast, all values of the dependent and independent variables are known. This uses a subsample of the data to fit the model and then compares its forecast against the known remaining values. While in the ex-ante or conditional forecast all the variables are not known with certainty, forecasts of the input variables are used to produce the corresponding forecasts of the dependent variable.

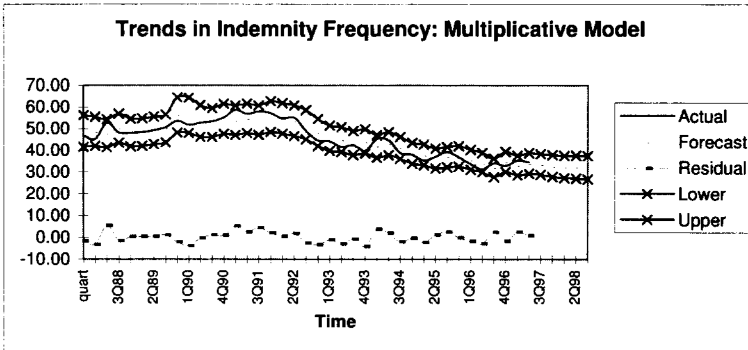
Note that we forecast the dependent variable(s) conditional on the assumptions of the independent variables. It is thus clear that any assumption about the input into the model affects the forecast generated.

As part of the rigorous process of model building and validation, we conducted an ex-post analysis of the models. The results are not discussed here but the general conclusion is that the nonlinear models have smaller root mean square errors of forecast. We discuss the results of the conditional forecasts in more detail. The plots below have the observed, the predicted, and the lower and upper confidence intervals of the frequency. The residuals are also provided which are found at the bottom of the graphs. Based on these plots, it seems quite apparent that, the nonlinear models have a much better fit and lower confidence bounds than the linear model. This is true in both the indemnity and medical cases.

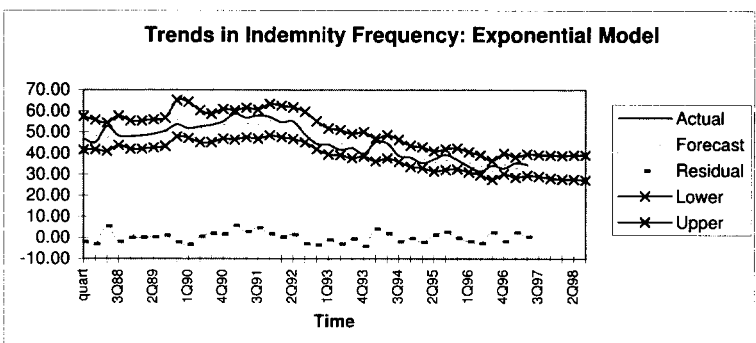
Also, the fits are much better in medical than in indemnity. Again, judging from the graphs, the nonlinear models are preferred to the linear specification. Note that the confidence bands are broader for the out of sample forecast reflecting the uncertainty in the model inputs. It appears that the linear model is much more sensitive to uncertainty in the input than the nonlinear ones. The interpretation of the confidence interval is that we are almost 95% confident that the realization of claims frequency for five quarters hence will fall within the confidence limits. As pointed out earlier, the real test of these models is when we get the data for the five quarters and compare them with the predictions for each of the models.

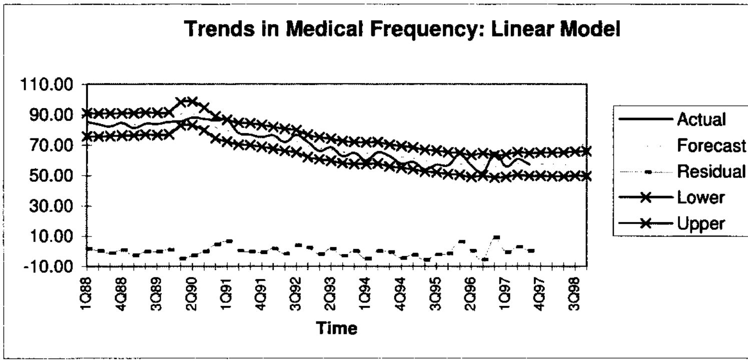
Graphs of actual frequency and predicted versus time in quarters.



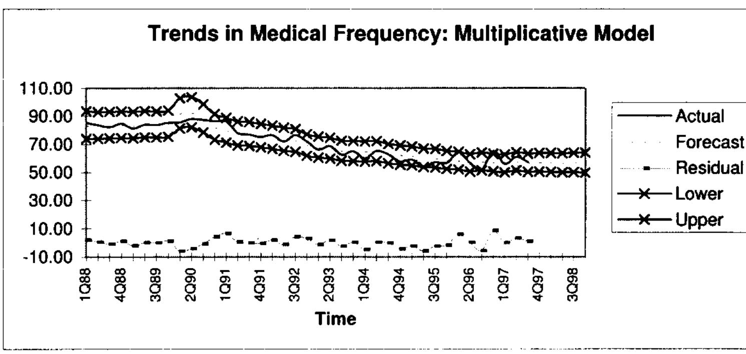


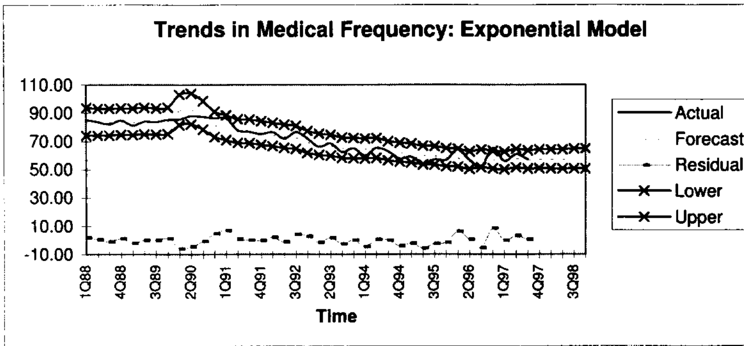
Graphs of actual frequency and predicted versus time in quarters.





Graphs of actual frequency and predicted versus time in quarters.





Notes on Trend Forecasting: - Actuarial versus Econometric Trends

This section discusses some issues related to frequency trends in insurance ratemaking process. Actuarial trending procedures employ what is known as deterministic trend models in estimating trends in frequency, severity and loss ratios. Deterministic trend models are often used in other areas such as economics, engineering and finance. Commonly used models in actuarial ratemaking are the linear and exponential time trend models. The econometric models developed are extensions to the trend models used by actuaries [see equations (1)-(3)]. In econometric parlance, the actuarial models are said to be nested in the econometric models. These equations reduce to the actuarial models when linear identifying restrictions, i.e., $\beta_2 = \beta_3 = 0$ are applied. This restriction is tested in all the models assuming that the null hypothesis is, H_0: the actuarial model is preferable to the econometric model. It is then possible to use the likelihood ratio statistic[7] which is approximately chi-square distributed

[7] The statistic is calculated as
$2(\log \textit{likelihood big} \bmod el - \log \textit{likelihood small} \bmod el)$.

with degrees of freedom equal to the number of restrictions imposed or the F-statistic to test the hypothesis. We employ both test statistics in this analysis. The tests have all been very highly significant leading to the rejection of the null hypothesis even at 1% level of significance. Hence, it can be concluded that using additional economic variables is worthwhile.

To answer some interesting practical questions, we use the data to compute some trend indications in two different ways. They will be labeled actuarial and econometric trends respectively. This is just for the purposes of taxonomy but not more. Note that there is no any good reason why, as far as we know, anyone of these procedures will be judged completely superior to the other. Much will depend on the intent and purpose of the analysis and who in fact conducts the analysis.

The results are reported for all three models and for both indemnity and medical claims frequency. Note that, actuarial trend models have been run and trend estimates are calculated to compare the actuarial trend forecasts and the econometric forecasts. The approach taken here is the actuarial methods. Notice that, while the two nonlinear models produce the percent growth rates directly, some actuarial or economic adjustments need to be made to the coefficient of the linear model in order to calculate the trend indication. The adjusted results will be reported for the linear models to ease comparison.

Tables 3 and 4 show the trend estimates using the Actuarial and Econometric models. Note that, the linear model estimates do not give the trend estimates prior to credibility. However, after some actuarial or economic adjustments, the slope coefficients can be converted into annual percentage changes. In economics, such adjustments include the calculation of the percent change at the mean or some other statistics. Here, an economic judgment is required by a trained and experienced individual, i.e. the judgment must be informed one. The growth rate at the mean is calculated as the trend coefficient of the linear model divided by the mean frequency over the entire series. The average values are respectively 45.5359 and 71.3026 for indemnity and medical frequency. Thus the growth rate is –0.0114[8] for the actuarial indemnity model. The results also include the standard errors of estimation which can be used to construct confidence intervals for the trend indication. In the case of the nonlinear models, the indications can be calculated by exponentiating the estimated coefficient of the trend variable. For example, using actuarial

[8] That is –0.5248 divided by 45.5359 equals linear trend. The other values can be calculated in a similar fashion.

estimate for indemnity will give an indication of 0.9878[9] while the econometric exponential model gives 0.9754. The econometric multiplicative model gives a slightly different number from the exponential, i.e. 0.9738. Similarly, this calculation can be done for the medical frequency models. Note that in estimating the coefficients, the number of observations is large compared to what actuaries would normally have available. Statistically, the more observations used, the smaller are the errors since they tend to cancel out. However, practical limitations and experience may warrant the use of new and recent observations.

---

[9] This is calculated as exp (-0.0122).

<u>Table 3 Trend Estimates</u>: Estimates from the linear and nonlinear actuarial models.

	Linear	*Multiplicative*	*Exponential*
Indemnity	-0.0114(0.0019)	--------	-0.0122(0.0018)
Medical	-0.0141 (0.001)	--------	-0.0142(0.00108)
IndRSquare	0.5115	--------	0.5489
MedRSquare	0.8438	--------	0.8364

Note that in Table 3, the multiplicative and the exponential estimates are the same in this case.

<u>Table 4 Trend Estimates</u>: Estimates from the linear and nonlinear econometric models.

	Linear	*Multiplicative*	*Exponential*
Indemnity	-0.0237(0.0035)	-0.0265(0.0028)	-0.0249(0.0031)
Medical	-0.0193 (0.0014)	-0.0204(0.0015)	-0.0203()
IndRSquare	0.6224	0.7682	0.7094
MedRsquare	0.9268	0.9243	0.9223

It is not surprising that the econometric models gave much better fit than the actuarial counterparts. In econometrics, it is expect that the bigger model will be at least as good as the smaller one but it is the magnitude of the gains that is dramatic in this case. As explained earlier these models use additional explanatory variables related to the economic demographic and social factors. The trends also show bigger declines in general than the actuarial model estimates. In actuarial trending procedures, several factors are taken into account. These include but not limited to goodness of fit measure, success of the model in making prior projection, etc. Thus a good model with a high explanatory power is welcome.

Summary and Conclusion

The study is conducted using the CWCI published quarterly data and three variants of econometric frequency models for both indemnity and medical experience. They all show that economic activity, i.e. the business cycle is still an important determinant of frequency. Most, if not all, previous studies are consistent with this observation. The important but not new message to practitioners and management is that economic activity is a significant determinant of claims frequency. Even in a declining frequency period, the decline may be slower than it would be in boom periods compared to periods of economic stagnation or recession.

Statistical tests have confirmed that nonlinearity is important in insurance claims frequency for the state of California Workers Compensation systems. This is because the two nonlinear models seem to outperform their linear counterpart. This makes sense because the real world itself is full of nonlinear relationships. In addition, it is shown that using economic variables resulted in a substantial payoff in terms of statistical performance. Trends have been calculated using actuarial and econometric models and the results have been discussed.

There are some theoretical issues regarding uncertainties in the input variables of an econometric model such as model uncertainty, variable and parameter uncertainty which have not been discussed in detail in this paper but are being considered as possible extensions to explore in future research.

Acknowledgments:

I thank Chun, Ruby, Tonya and Ahmed Ustarz for their help. I also thank Ziv for interesting and simulating discussions.

References

Kahley, W. J. California Workers' Compensation Claims Frequency Forecast, California Workers Compensation Institute, 2000.
Pindyck, R.S. and Rubinfeld, D.L., 1997, 4th Edition.
Ussif, A.M., An international analysis of workplace injuries, 2002.