

Predicting Future Paid Losses

A Major Qualifying Project Report
Submitted to the Faculty of the
WORCESTER POLYTECHNIC INSTITUTE
in partial fulfillment of the requirements for the
Degree of Bachelor of Science
in Actuarial Mathematics

by:

Haiying Liu

Aryeh Shatz

Mingzhu Zheng

Submitted:

Sponsor: Hanover Insurance Group

Approved by:

Professor Jonathan Abraham, Major Advisor

Abstract

This project seeks the best mathematical model to forecast the future losses for each state from a given series of historical data by Hanover Insurance Group. There are nine models chosen after analyzing all the states' data. The score method is applied to seek out the accurate models among the nine models. Comprehensive analysis also shows that some external experience needs to be used in order to make the prediction more precise.

Acknowledgements

Our group would like to thank the following individuals and groups for their time, effort, support, and guidance in completing this project:

From Hanover Insurance:

Alyssa Lopes

Jonathan Blake

From Worcester Polytechnic Institute:

Jon P. Abraham

Norman Lam

Ethan Brown

Executive Summary

The Hanover Insurance Group is a Worcester-based insurance company offering a variety of insurance products. The company uses its historical data to evaluate trends in its insurance policies using several internal methods.

The goal of this project was to **identify and evaluate historical trends** in frequency, severity, and pure premium, and to use this information to predict future loss experience.

Steps included:

- Testing historical data
- Researching alternative and innovative loss trend selection methods
- Fitting the loss trend methods with the real data to evaluate the accuracy of each method
- Recommending the best method for future loss trend prediction

Three categories of methods were used for prediction.

1. The first involve standard plots of data against time and fitting a curve through the points. These trend line methods include Linear, Exponential, Power, and Logarithm fits.
2. The next category focused on the rate of change between two consecutive data points, plotting future data based on the previous few data points. These methods are the Level, Quarterly, and the Yearly Methods.
3. The final category was prediction based on autocorrelation. Two methods were developed here: the Auto-regressive and Linear Exponential Models.

Two scoring methods were used to compare the accuracy of each method. The residual method compared the sum of the absolute value of “actual minus expected” for each data point.

The margin of error method divided the residual error by the average value of the actual data, resulting in a percentage error.

The purpose of this project was to develop a general model by which our sponsor could accurately predict future losses. The mathematical models provide a reasonably accurate approach to predicting future loss experience under many scenarios; however, testing on a larger volume of data and over a longer period of time will be needed to make sure the models retain their consistency and accuracy. Further, these same approaches could potentially be used on other blocks of business.

Table of Contents

Abstract	2
Acknowledgements.....	3
Executive Summary	4
Table of Contents	6
Table of Tables	8
Table of Figures	9
1 Introduction.....	10
1.1 Basic Introduction about Hanover	10
1.2 Problem Statement	10
2 Background.....	12
2.1 Basic Terminology.....	12
2.2 Attributes	16
3 Methodology.....	19
3.1 Basic (Linear, EXPOENTIAL, Power, log).....	20
3.2 Rate of change (Quarterly, Yearly, Level).....	22
3.3 Auto-Correlation model	24
3.3.1 Auto-Correlation coefficient (r).....	24
3.3.2 Auto-Regressive model.....	26
3.3.3 Linear Exponential model	27
4 Score Methods.....	29
4.1 Data Organization	29
4.2 Data Automation.....	29
4.3 Average Residual	31
4.4 Margin of Error.....	31
5 Analysis	33
5.1 Colorful tables	33
5.1.1 Average Residual.....	33
5.1.2 Margin of Error.....	34
5.2 Best Prediction Models	34
5.2.1 Model Testing.....	34

5.2.2	Test Results	36
6	Conclusion / Recommendation	39
	Appendix A: NY_PIP with Data range 1-20	40
	Appendix B: Each State's % of total losses	41
	Appendix C: Detailed Model Test	45
	Reference	47

Table of Tables

Table 3-1: Shifted Data	26
Table 4-1: Average Residual	31
Table 4-2: Margin of error.....	32
Table 5-1: Margin of Error Table for NY_Severity with data range start 10 length 30	36
Table 5-2: Residual Table for NY_Severity with data range start 10 length 30	36
Table 5-3: Best Model for MA, MI, NY, NJ	36

Table of Figures

Figure 3-1: Linear Model	20
Figure 3-2: Exponential Model.....	21
Figure 3-3: Power Model	21
Figure 3-4: Log Model.....	22
Figure 3-5: Linear Exponential Coefficients	27

1 Introduction

1.1 Basic Introduction about Hanover

Based in Worcester, Massachusetts, The Hanover Insurance Group, founded in 1852 as a property and casualty insurance company, is one of the 500 largest publicly-traded companies in the United States. It is the parent company of two divisions, Hanover Insurance and Citizens Insurance. Both divisions serve customers with auto, home, and business insurance.

Having been focusing on Property and Casualty products for nearly two centuries, the Hanover Insurance Group has managed through periods of economic prosperity and adversity. It has provided various insurance protections to millions of Americans throughout the country. Citizens Insurance was regarded as the first automobile insurer in the state of Michigan, was acquired by the parent company in 1974.

Today, the Hanover Insurance has grown to provide a wide range of Personal Lines and Commercial Lines products to meet customers' need. The products include: Business Owner's Policy, Automobile, Commercial Auto, Commercial Package, Home, Renter, Condominium and Dwelling Fire, Workers' Compensation, Umbrella, Inland Marine, Boat, Bond, and Specialty.

1.2 Problem Statement

To maintain favorable annual growth and success as a major insurer, the Hanover Group places a strong emphasis on understanding the company's loss trend and history to control risks and

predict future losses. Specifically, for this project, the process of loss control depends on a series of algorithms that analyze and compare the performance of future loss predictions. In order to satisfy the increased business demands and to have an accurate future loss expectation, we focus on establishing methods in order to recognize the best loss trend prediction performance.

2 Background

2.1 Basic Terminology

In order to understand trends in loss frequency and severity, we completed a comprehensive evaluation project to research the accuracy of past Auto and Homeowners selected trends based on the Hanover Insurance Group's losses history data. This project consisted of

1. Testing historical data
2. Researching alternative and innovative loss trend selection methods
3. Fitting the loss trend methods with the real data to evaluate the accuracy of each method
4. Recommending the best method for future loss trend prediction

We proposed to examine the accuracy of the current trend selection methods to prepare for research on an alternative and innovative loss trend selection method. The new trend selection method will be able to accurately predict future loss experience.

Currently, the Hanover Insurance Group is using their own methods for predicting to evaluate their insurance policies. Each insurance policy is serving customers in difference states with auto insurance coverage.

Auto insurance coverage is packaged into six different coverage types. The customers determined what they were required to purchase and what needed to be protected. That is, one can buy insurance in the event that they caused damage to their own property and can also buy insurance in the case where someone else damaged their property. A breakdown explanation of what each insurance coverage type protects is needed in order to understand our project.

Bodily injury liability, BI, covers other people's bodily injuries or death for which the insurance policyholder is responsible. This policy does not cover vehicles. Bodily injury coverage is mandatory in most states. It provides a legal defense in the case where another party in the accident files a lawsuit against this customer. Claims for bodily injury include medical bills, loss of income, pain, and suffering. Usually, in the event of a serious accident, the insurance policyholder wants enough insurance to cover a verdict against her in a lawsuit, without jeopardizing her personal assets. The dual coverage limits refer to the maximum amount that will be paid per person per incident. It is, therefore, wise for the customers not to select coverage limits that are too low; if the accident damages exceed their limits, they will be held responsible for the amount above their coverage limits.

Property damage liability, PD, covers the customer whose car damages someone else's property. Usually the claim object is a car, but it could be any property damaged in an accident, for example, a house or a fence. The coverage limits refer to the maximum amounts that will be paid per accident, and coverage is limited to the terms and conditions contained in the policy. It is, again, generally wise for customers to purchase enough of this insurance to cover the amount of damage their car might inflict. In the state of Michigan, there exists a coverage that is related to property damage liability – limited property damage. It provides protection if the insurance policyholder is at fault in an accident that causes damage to another vehicle.

Physical damage coverage covers the customer's vehicle. In limited scenarios it covers other vehicles that one may be driving for losses resulting from incidents other than collision. There are two types of physical damage coverage: comprehensive coverage and collision coverage. Comprehensive insurance, CM, covers damage to the customer's car if it is stolen or damaged by flood, fire, or animals. The amount of coverage provided typically refers to the portion of a claim the customer is responsible for paying. This is also known as the deductible. Those whose cars are either financed, leased, who have a newer vehicle, or one in excellent condition will benefit the most from buying comprehensive coverage. One who has an older car or one in poor condition may, however, not want to pay for this coverage. Sometimes if customers want to keep their premiums low, they select as high a deductible as they feel comfortable paying out of their pockets. Comprehensive physical damage is not required by a state, but if a person has a loan or a lease then the holder will require it.

Another type of physical damage coverage is collision coverage (CO). It covers damage to the car of the insurance policyholder when the car hits, or is hit by another vehicle or object. The coverage pays to fix the customers' vehicle minus the deductible they choose. It is usually recommended for customers who have older cars to consider dropping this coverage, since collision coverage is normally limited to the cash value of their own car. Like comprehensive physical damage, this coverage while not required by state, someone with a loan or a lease will need it.

Personal Injury Protection coverage, PIP, covers within the specified limits, the medical, hospital, funeral expenses of the insured, others in his vehicles, and pedestrians struck by him. Usually it benefits the policyholder, the policyholder's relatives in the same household, and passengers. In some states, it protects the policyholder, family members who are injured while riding in someone else's car, or pedestrians struck by another vehicle. It is only available in certain states. Total payments covered by Personal Injury Protection are the maximum amounts that will be paid per person for any combination of covered expenses (some states offer limits and others set it to an amount like \$10,000). Specific limits and coverage vary by state. Depending on the state, the covered parties below and the amount of protection may vary. It is recommended for people who don't have health insurance that adequately covers the expenses listed above or people who carpool or frequently drive with passengers to have personal injury protection coverage.

Combined Single Limit coverage, CSL, combines both bodily injury liability and property damage liability insurance under a single limit. The insurance company will pay up to the stated limit on a third party claim regardless of whether the claim was for bodily injury, property damage, or both. For those who lease a car, this coverage is not always required by the state, but may be mandated by a specific leasing company. Usually when financing a car, whether a lease or loan, one will normally be required to have not only the state required liability coverage on the vehicle but also physical damage coverage of collision and comprehensive. If the leasing company requires Combined Single Limit then this would mean combining your liability limits instead of the normal split limits. When the policy holder makes a claim, the limit for the CSL is

the total that the insurance provider will pay for all bodily injuries and property damage caused in one accident. Whatever the number of people injured or the portion of bodily injury or property damage is, CSL will cover it.

2.2 Attributes

The availability of different forms of coverage varies from insurance company to insurance company in each state. In our project, we studied the distribution of the coverage in each state in order to get a hold of the loss trend. By doing so, we analyzed the trend of specific insurance attributes: severity, frequency and pure premium.

These three attributes were derived from the paid loss amount, earned exposure, and number of claims. Exposure represents the number of people insured by the insurance company. It is the basis to which rates are applied to determine premium. Exposures may be measured by payroll, as in workers compensation or general liability, receipts, sales, square footage, area, or man-hours for general liability. In automobile it can be measured by unit while in property insurance it can be measured by per \$1,000 of value. Claims represent the number of people claiming a loss when accident happens.

Severity is the amount of damage inflicted by a single loss. Severity was calculated by dividing the paid losses by claims, seen as below:

$$\textit{Severity} = \textit{Losses/Claims}$$

Frequency is the likelihood that a loss will occur; it is calculated by dividing the claims by the exposure:

$$\textit{Frequency} = \textit{Claims/Exposure}$$

It is expressed as low, moderate, and high frequency. Low frequency refers to losses which have rarely happened in the past and is not likely to occur in the future. Moderate frequency means that the loss event has happened once in a while and can be expected to occur sometime in the future. High frequency is the loss event happens regularly and can be expected to occur regularly in the future. Usually, in the auto industry, the compensation losses normally have a high frequency as do automobile collision losses. General liability losses are usually of a moderate frequency, and property losses often have a low frequency.

Pure premium is the part of the premium which is sufficient to pay losses and loss adjustment expenses, but not other expenses. It is also called “loss cost”, the actual or expected cost to an insurer of indemnity payments and allocated loss adjustment expenses. Pure premium was calculated by dividing the losses by the exposure:

$$\textit{Pure Premium} = \textit{Losses/Exposure}$$

Pure premium does not include overhead costs or profit loadings. Historical loss costs reflect only the costs and allocated loss adjustment expenses associated with past claims. Prospective loss costs are estimates of future loss costs, which are derived by trending and developing

historical loss costs. Insurers add their own expense and profit loadings to these loss costs to develop rates which are then filed with regulators. The pure premium is developed by dividing losses by exposure, disregarding any loading for commission, taxes and expenses.

3 Methodology

Without an effective trend selection method to help predict future loss experience, the company will suffer from losses. An accurate approach to evaluate trends in loss, frequency, and severity is needed. An effective trend selection method may help the company to have a better understanding of the loss history, prevent an escalated loss in profit, and promote competitive service and premium rates in the market.

Many people, including the Hanover Insurance Group, have been studying frequency, severity, and pure premium trends to find accurate ways to predict future losses. There are three categories of methods used for prediction. The first is to plot the data against time and plot a curve through the points. The methods with this attributes include the linear, exponential, power, and log. These four basic models underlie many of the statistical analyses that are used in applied science and social research. They are the foundation for many social and mathematical analysis including factor analysis, cluster analysis, multidimensional scaling, and others. Because of their generality, the model is important for people who conduct statistical research and find relationships among variables. The next category deals with plotting future data based on the previous few data points. These methods are the quarterly, repeat, and the yearly. The final category is predicting based on autocorrelation. Autocorrelation describes the correlation between the original data set and the same data set shifted forward.

The data that was used as the example of the models was NY PIP severity from quarters 1-20. (See Appendix A)

3.1 Basic (Linear, Exponential, Power, Logarithmic)

Linear

By plotting the data using a linear function, each data point was found using the equation $x=at+b$. “t” was the time of the data, “a” was its slope, and “b” was the y-intercept. For example, in the case provided, “a” was found to be 152.2743 and b was 2412.0896. Therefore when predicting quarters 21-24, simply plug in the appropriate quarter number into “t” and keep a and b the same. Hence for the 21st quarter, the severity was 5610.

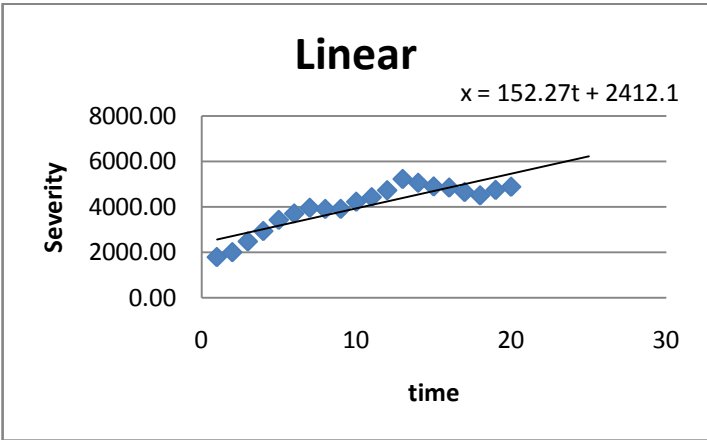


Figure 3-1: Linear Model

Exponential

The exponential method was just like the linear except the data was plotted using the equation $x = a * e^{bt}$. If the data had an exponential distribution, then as time passes the data would have increased or decreased more rapidly as time increased. Going back to the NY severity data, a was 2413.33 and b was equal to 0.04. Therefore at the 23rd quarter, the severity was equal to 6727.

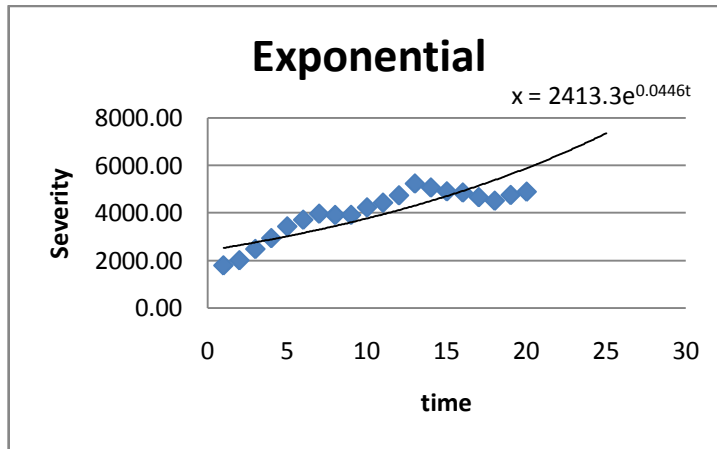


Figure 3-2: Exponential Model

Power

In the Power method, the predicted values increased or decreased at a slower rate as the quarters increased. In other words, the exponential curve will have a positive second derivative while the logarithmic will be negative. The equation used to plot the data was $x = a * t^b$. Again using the NY severity data, a and b were 1759.08 and .37 respectively. Therefore the severity value at the 22nd quarter was equal to 5528 5295.

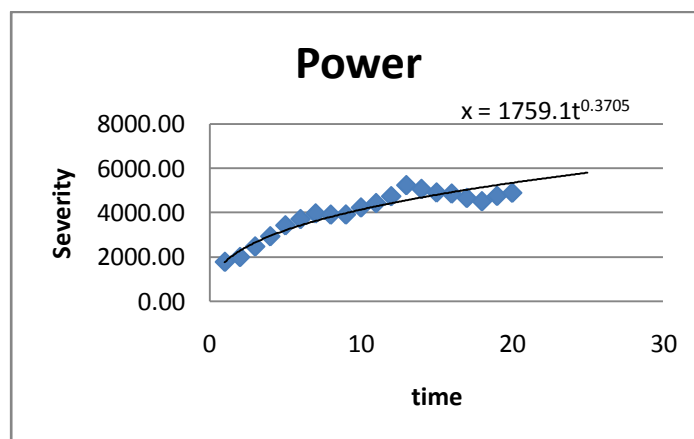


Figure 3-3: Power Model

Logarithmic

The logarithmic method is simply the logarithmic equation applied to the power model and therefore has a similar curve. The data was plotted using the equation $y = a \cdot \ln(t) + b$. Returning to the NY severity example where a and b were equal to 1210.36 and 1448.91, respectively, the predicted severity value for the 24th quarter was 5295.

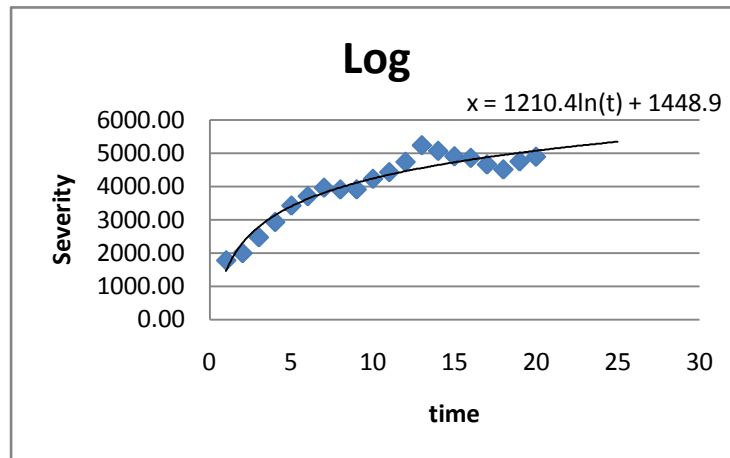


Figure 3-4: Log Model

3.2 Rate of change (Quarterly, Yearly, Level)

Quarterly

The quarterly method was like the exponential method but used a slightly different assumption. The quarterly method always assumed that the rate of change between the current quarter and the next one was directly based on the rate of change between the previous quarter and the current one. The method's equation was

$$x_t = (x_{t-1}/x_{t-2}) * (x_{t-1})$$

In the example of NY PIP severity (see Appendix A), the value of the 20th quarter was 4884.63. the predicted value of severity in the 21st quarter was 5029. To predict the 22nd quarter, one would have used the rate of change between the predicted 21st and the actual 20th. Therefore, the

22nd quarter was equal to 5178:

$$\frac{5029}{4884.63} * 5029 = 5178$$

Yearly

The Yearly method was just like the quarterly method except it assumed that the rate of change between the current quarter and the next one was related to the rate of change between the severity of four quarters ago and the severity of five quarters ago. That is, the rate of change between two quarters is related to the rate of change between those same two quarters in the previous year. Hence the equation was

$$x_t = (x_{t-4}/x_{t-5}) * (x_{t-1}).$$

Again, using the same NY severity example (see Appendix A). The value of the 16th quarter was 4848.63, and the 17th quarter was 4651.72, since the value of the 20th quarter was 4884.63, then the predicted value of the 21nd quarter was 4686:

$$\frac{4651.72}{4848.63} * 4884.63 = 4686$$

Level

The Level method used an almost entirely different assumption than all the other methods for predictions—it assumed there was no trend between the data. The equation for this method was simple which is $x_t = x_{t-1}$. That is, there was no way of knowing whether the value of the next quarter will increase, decrease, or stay the same. Therefore the only approach was to guess that it will be approximately the same value as the current quarter. In the example of NY severity, the

predicted severity of all the quarters, 21-24, was equal to the severity at the 20th quarter which was 4885.

3.3 Auto-Correlation model

3.3.1 Auto-Correlation coefficient (r)

Auto-Correlation, the key factor in building the Auto-Regressive and Linear Exponential models, can be derived from Pearson's Correlation. Pearson's correlation coefficient measures the degree to which two different sets of variables are linearly related. It could be any number between negative one and one. A positive correlation indicates a same linearly direction between two variables. That is, as the value of one variable increases, the values of another variable also increases too. A correlation coefficient of exactly +1 indicates a perfect positive fit. Visa versa, a negative correlation indicates an opposite linearly direction between two variables where as one value increases, the other one would decrease. The closer the absolute value of the correlation coefficient to one, the stronger the linear relationship appears between two variables.

Denoted as r, the mathematical formula for computing the Pearson correlation coefficient is:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)S_x S_y}$$

n is the number of pairs of data, x and y are the two different set variables where the correlation may exist, and S_x and S_y are the standard deviations for X and Y respectively. The auto-correlation coefficient works basically the same way as the Pearson's correlation coefficient except there is only one sample data. Since correlation required two sample data sets, a new set

was created by shifting the original data. X_{t+k} was the original sample shifted forward k spaces. K represented the order of the shifted data. Table 3-1 demonstrated how to get k_{th} order data with $k=1$ and $k=2$. X_t is the original data set. The first order shifted data X_{t+1} moved X_t column up by one space, the second order shifted data X_{t+2} move X_t column up by two spaces. Then find out the 1st and 2nd auto-correlation for the pairs (X_t, X_{t+1}) and (X_t, X_{t+2}) respectively.

The auto-correlation coefficient formula is:

$$r_k = \frac{\sum_{t=1}^{n-k} (x_t - \mu)(x_{t+k} - \mu)}{\sum_{t=1}^n (x_t - \mu)^2}$$

μ is the only mean of the original sample since there is just one sample given. Since the two standard deviations are now equal, the denominator is simply the original sample variance multiplied by “ $n-1$ ”.

Applying the derived auto-correlation coefficient from Pearson’s correlation to the example we got:

$$u = (1772.67 + 1989.64 + \dots + 4884.63) / 20 = 3208.776$$

$$r_1 = [(1772.67 - u) * (1989.64 - u) + \dots + (4744.14 - u) * (4884.63)] / [(1772.67 - u)^2 + \dots + (4884.63)^2] = 0.7758$$

$$r_2 = [1772.67 - u) * (2466.11 - u) + \dots + (4503.05 - u) * (4884.63)] / [(1772.67 - u)^2 + \dots + (4884.63)^2] = 0.7321$$

t	X_t	X_{t+1}	X_{t+2}
1	1772.67	1989.64	2466.11
2	1989.64	2466.11	2924.42
3	2466.11	2924.42	3417.84
4	2924.42	3417.84	3699.57

5	3417.84	3699.57	3954.49
6	3699.57	3954.49	3900.85
7	3954.49	3900.85	3904.89
8	3900.85	3904.89	4221.17
9	3904.89	4221.17	4421.22
10	4221.17	4421.22	4729.64
11	4421.22	4729.64	5228.22
12	4729.64	5228.22	5059.26
13	5228.22	5059.26	4897.24
14	5059.26	4897.24	4848.63
15	4897.24	4848.63	4651.72
16	4848.63	4651.72	4503.05
17	4651.72	4503.05	4744.14
18	4503.05	4744.14	4884.63
19	4744.14	4884.63	N/A
20	4884.63	N/A	N/A

Table 3-1: Shifted Data

3.3.2 Auto-Regressive model

Let X_t be the predicted variable, the Auto-regression model uses the correlation coefficient to relate X_t to the immediately past value with a First Order auto-correlation coefficient. The equation to describe this relation is:

$$X_t = r_1 * X_{t-1} + (1 - r_1) * \text{Mean}(X_{t-1}, \dots, X_1).$$

In other words, the current term of the series can be estimated by a linear weighted sum of previous terms in the series. The weights are the correlation coefficients (which always add up to 1). For the Auto-Regression model, when the two sets of data maintain a strong correlation, then the predicting value is very close to the immediate past value. Otherwise, the predicting value will lean towards the mean of all past observations.

3.3.3 Linear Exponential model

Linear Exponential model shares the same idea in the correlation process as the Auto-regression model, except the Linear exponential model not only directly correlates the immediately past value, but takes all past values in to consideration and assigns weights on each past observation.

The equation to describe this process is denoted as:

$$X_t = r_1^0(1 - r_1)X_{t-1} + r_1^1(1 - r_1)X_{t-2} + r_1^2(1 - r_1)X_{t-3} + r_1^3(1 - r_1)X_{t-4} + \dots$$

The model works in such a way that although based on all past observations, the weights are not all equal to one. Instead, the weights applied to recent observations are larger than the weights applied to earlier observations. This weighting structure works well because it is reasonable to consider the recent data plays a more important role in predicting the next data than the older data. That is, the weights follow a negative exponential curve, hence the name “Linear Exponential” where as the correlation coefficient increases, the flatter the slope (see Figure 3-5). Also, by adding all the weights together, it forms a geometric series which converges to one. It is therefore unnecessary to divide all the weights by the sum of the coefficients.

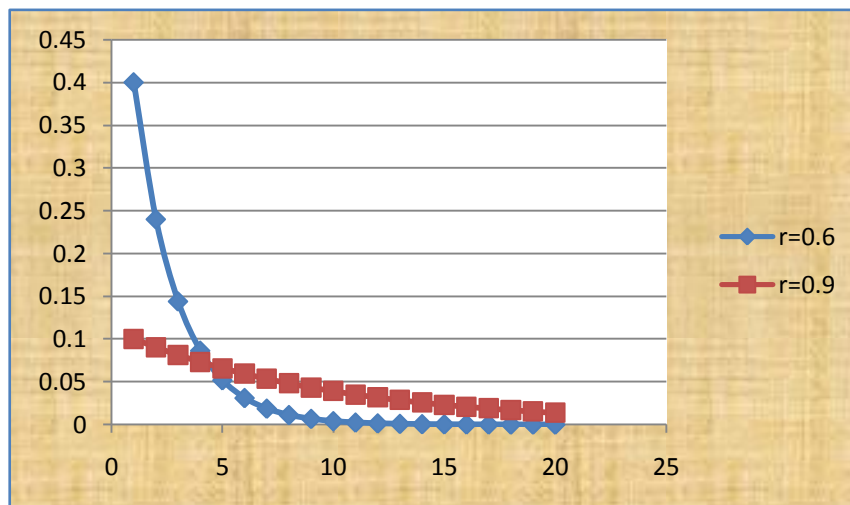


Figure 3-5: Linear Exponential Coefficients

4 Score Methods

4.1 Data Organization

The data were split up into twenty-three different states, six different coverage, and three different attributes. For each data set, there were forty-eight points representing each quarter of data. The first quarter began at 09/96 and the last quarter was at 06/08. Even though the data were split quarterly over each year, the number represented the entire data for the previous 12 months. For example, the data at quarter 6/05 included all the data ranging from 6/04 to 6/05. Likewise, the data at quarter 9/05 contained all the data from 9/04 to 9/05. Each data set was labeled from 1 to 48, 1 represented the date 09/96, 2 represented the date 12/96, and so on so forth, 48 represents the last date 06/08.

4.2 Data Automation

Automatic tool (excel with macro) was created to forecast the future losses by using all of the nine models (four basic trend line models, three rate of change models, two auto-correlation models). On the "Selection" excel spreadsheet, there were two main parts: the first part asked the user to select the specific state, coverage, and attribute that he focused on. The second part asked the user to select the range of the data among the forty-eight data points to predict the following four quarters, and then compared the prediction value with the actual value. "Start Box" meant the first point in the data range. "Length Box" means the number of data points used in the data range. The end of the data point and the prediction period (the following four quarters) automatically appeared once this two were chosen. For example, if the user wanted to

see the result of state (NY), coverage (BI), and attribute (Severity) for the prediction period 33-36 with the data range 4-32, the information is represented in the following image:

Please select the state, coverage and attribute	
State	NY
Coverage	BI
Attribute	Severity
Please select the data range to predict next 4 quarters	
Start	4
Length	29
End	32
Prediction Period	33-36

Fill AvgResidualTable & MarginErrTable

Clear AvgResidualTable & MarginErrTable

Note: the latest prediction period is 45-48

There were two buttons on the “Selection” excel spreadsheet. When the user clicked the upper button, the residual table and the margin of error table were filled. These two tables showed the result for all of the attributes and all of the coverage for one specific state and data range, instead of one specific attribute with respect of coverage. The user could ignore the coverage and attribute when he wanted to fill the both tables if he knew the state and the data range which interested him. When the lower button was clicked, the contents in these two tables would disappear.

The Data Automation tool was attached in the CD.

4.3 Average Residual

Residual is the difference between the sample data and the data that is generated from the fitted function. Let x be the sample data and x' be the fitted data that was generated from the prediction methods. The residual would be defined as: $r = |x - x'|$. Four of the prediction quarters had the same importance, we weighted each quarter evenly. The average residual was always the mean of the four consecutive residuals, and calculating as the following:

$$\bar{r} = (r_1 + r_2 + r_3 + r_4)/4$$

Table 4.1 shows an example of how the average residual was calculated using New York PIP coverage and pure premium attribute with the Linear model using data range 1-25 to predict 26-29.

Actual Data	Prediction	Residual	Average Residual
85	94	9	$(9 + 2 + 2 + 4)/4 = 4.25$
94	96	2	
96	98	2	
97	101	4	

Table 4-1: Average Residual

4.4 Margin of Error

The margin of error showed the accuracy of the prediction in terms of the ratio of the average residual to the average of the actual value. It's calculated using the formula:

$$m = \frac{\bar{r}}{\bar{x}}$$

Denoting margin of error as m , the average residual as \bar{r} , and the mean of the actual values that were trying to be predicted as \bar{x} .

Table 4.2 shows an example of how the average residual was calculated using New York PIP coverage and pure premium attribute with the Linear model using data range 1-25 to predict 26-29.

Actual Data	Prediction	Residual	Average Residual \bar{r}
85	94	9	$(9+2+2+4) / 4 = 4.25$
94	96	2	
96	98	2	Mean of Actual Data \bar{x}
97	101	4	$(85 + 94 + 96 + 97) / 4 = 93.10$

Table 4-2: Margin of error

the margin of error is: $m = \frac{4.25}{93.10} = 4.56\%$

5 Analysis

5.1 Colorful tables

Average residual tables and margin of error tables were constructed-one for each attribute-for each state in a certain prediction period. Each Table included six coverage and nine prediction models. Then we assigned colors to each cell.

5.1.1 Average Residual

In the average residual table, each row was treated independently from the other rows. Colors were assigned based on the value of the average residual in comparison with the other average residuals in the same row. In the order of the lowest average residual to the highest, the respective colors were dark green, light green, yellow, orange, and red.

In the New York severity Average Residual table below, under the coverage BI row, predicted data derived from the Logarithmic model was very far off from the sample data in three instances. It was labeled as red, indicating this model as quite inaccurate.

Coverage	Data Period	Prediction Period	Linear 30	Exponential 30	Log 30	Power 30	Yearly	Level	Quarterly	Auto-Regressive	Linear Exponential
BI	12-1998 to 9-2003	6-2006 to 3-2007	6544	6394	13202	9687	8522	2470	4409	6735	6003
CM	12-1998 to 9-2003	6-2006 to 3-2007	57	54	30	31	24	26	26	18	30
CO	12-1998 to 9-2003	6-2006 to 3-2007	217	189	701	416	93	252	304	368	539
CSL	12-1998 to 9-2003	6-2006 to 3-2007	1566	1732	1673	1286	2515	1450	792	871	1233
PD	12-1998 to 9-2003	6-2006 to 3-2007	76	88	388	116	280	81	293	63	169
PIP	12-1998 to 9-2003	6-2006 to 3-2007	947	1028	602	243	288	272	505	116	58

5.1.2 Margin of Error

Unlike the average residual table, each cell in the margin of error table was independent of all other cells regardless of the row and column. A margin of error value was green for less than five percent, light green between five and ten percent, yellow between ten and thirty percent, orange between thirty and fifty percent, and red above fifty percent. The NY Severity 30 period margin of error table below, sometimes there were no inaccurate methods since none of the cells were filled with a red color. It was also easier to conclude which coverage was hard to predict.

Coverage	Data Period	Prediction Period	Linear 30	Exponential 30	Log 30	Power 30	Yearly	Level	Quarterly	Auto-Regressive	Exponential Smoothing
BI	12-1998 to 9-2003	6-2006 to 3-2007	17.20%	16.81%	34.70%	25.46%	22.40%	6.49%	11.59%	17.70%	15.78%
CM	12-1998 to 9-2003	6-2006 to 3-2007	7.18%	6.91%	3.83%	3.90%	3.04%	3.33%	3.29%	2.30%	3.76%
CO	12-1998 to 9-2003	6-2006 to 3-2007	5.84%	5.08%	18.83%	11.17%	2.51%	6.77%	8.16%	9.90%	14.48%
CSL	12-1998 to 9-2003	6-2006 to 3-2007	16.89%	18.69%	18.04%	13.88%	27.13%	15.65%	8.54%	9.39%	13.30%
PD	12-1998 to 9-2003	6-2006 to 3-2007	2.79%	3.22%	14.26%	4.27%	10.28%	2.98%	10.78%	2.32%	6.21%
PIP	12-1998 to 9-2003	6-2006 to 3-2007	15.97%	17.33%	10.14%	4.09%	4.85%	4.59%	8.52%	1.95%	0.98%

5.2 Best Prediction Models

5.2.1 Model Testing

Hanover had 80% of business in the four states: MI, MA, NJ, and NY (See Appendix A). Our main goal was to analyze these four states, and then find out the best models for them. This result would be assumed to apply to the rest of the states as well.

When analyzing the tables, if any of the columns were mainly green, then that prediction (which was represented by that particular column) was the best no matter what coverage. It was more important that this was true for the margin of error tables rather than the average residual comparison tables. That is, it was more significant that a method be consistently accurate rather than it being consistently the best method. Many times, a certain method was green in the average residual table and many of the other methods were green in that same section of the

margin of error table. In such a case, all the methods that were green in the margin of error table were good predictions.

While the prediction that was green in the average residual table was technically the best prediction, the amount that it was best by was little enough to conclude that it may not really be the best method overall. The most conclusive result would be if a certain method is the only one to be green in both tables. That particular method would then definitively have been the best method.

To choose the best models, we set up some rules. First we looked at the colors of the margin of error table, and considered the dark green and light green by comparing column to column for each attribute (each column consisted of six coverages). The best models would have had the most of those two colors. If any of the models had the same amount of dark green and light green, we would then check the residual table for whichever had the most amount of greens in order to find the best model. If the models still had the same amount of green, we considered the models equal. If any model included at least two red colors, then we would not consider this model since it meant that the model created some outliers. For example, if we wanted to know which model was the best for NY_Severity with data range 10-39, we first examined margin of error table. Table 5-1 obviously showed the Level and Auto-Regressive models stood out, because both of them had a total of four dark and light greens. To determine which model of the two was better, we used the average residual table. Table 5-2 showed that Auto-Regressive had three dark greens while level only had one. We concluded that Auto-Regressive was the best model for NY_Severity with data range 10-39.

Coverage	Data Period	Prediction Period	Linear 30	Exponential 30	Log 30	Power 30	Yearly	Level	Quarterly	Auto-Regressive	Linear Exponential
BI	12-1998 to 9-2003	6-2006 to 3-2007	17.20%	16.81%	34.70%	25.46%	22.40%	6.49%	11.59%	17.70%	15.78%
CM	12-1998 to 9-2003	6-2006 to 3-2007	7.18%	6.91%	3.83%	3.90%	3.04%	3.33%	3.29%	2.30%	3.76%
CO	12-1998 to 9-2003	6-2006 to 3-2007	5.84%	5.08%	18.83%	11.17%	2.51%	6.77%	8.16%	9.90%	14.48%
CSL	12-1998 to 9-2003	6-2006 to 3-2007	16.89%	18.69%	18.04%	13.88%	27.13%	15.65%	8.54%	9.39%	13.30%
PD	12-1998 to 9-2003	6-2006 to 3-2007	2.79%	3.22%	14.26%	4.27%	10.28%	2.98%	10.78%	2.32%	6.21%
PIP	12-1998 to 9-2003	6-2006 to 3-2007	15.97%	17.33%	10.14%	4.09%	4.85%	4.59%	8.52%	1.95%	0.98%

Table 5-1: Margin of Error Table for NY_Severity with data range start 10 length 30

Coverage	Data Period	Prediction Period	Linear 30	Exponential 1 30	Log 30	Power 30	Yearly	Level	Quarterly	Auto-Regressive	Linear Exponential
BI	12-1998 to 9-2003	6-2006 to 3-2007	6,544	6,394	13,202	9,687	8,522	2,470	4,409	6,735	6,003
CM	12-1998 to 9-2003	6-2006 to 3-2007	57	54	30	31	24	26	26	18	30
CO	12-1998 to 9-2003	6-2006 to 3-2007	217	189	701	416	93	252	304	368	539
CSL	12-1998 to 9-2003	6-2006 to 3-2007	1,566	1,732	1,673	1,286	2,515	1,450	792	871	1,233
PD	12-1998 to 9-2003	6-2006 to 3-2007	76	88	388	116	280	81	293	63	169
PIP	12-1998 to 9-2003	6-2006 to 3-2007	947	1,028	602	243	288	272	505	116	58

Table 5-2: Residual Table for NY_Severity with data range start 10 length 30

We chose seven different data ranges to test the nine models. The details of the model test were in Appendix B. The final result (best model) listed in the following tables:

	MA	MI	NY	NJ
Frequency	Level	Auto	Auto	Level
Severity	Auto	Level	Level	Auto
PurePremium	Level	Auto	Level	Level

Table 5-3: Best Model for MA, MI, NY, NJ

5.2.2 Test Results

After analyzing myriads of tables and data sets, the most consistent models were the Auto-Regressive and the Level. The Auto-Regressive model was especially accurate with a strong auto-correlation coefficient (greater than or equal to .8). It was, therefore, imperative to use the correct amount of data to give the model the highest auto-correlation coefficient. While this number varied in the tested data, thirty data points was most often the best choice. However, as

Hanover collects more and more data, that number may still change. The level method, on the other hand, was consistent throughout. Both of these methods have a very similar premise. The Level model always assumed the coefficient multiplied by the last known quarter is equal to 1. With a high auto-correlation coefficient, the coefficient multiplied by the last known quarter is close to one. Hence both of these models assumed that the first predicted quarter was extremely close to the last known quarter.

The other somewhat consistent prediction models were the rate of change models. The quarterly was successful when the data trend was monotonic. That is, if the data was increasing and kept increasing during the prediction period—or vice versa—then the quarterly method would be the most accurate. The Yearly method was precise when the trend in the data was seasonal.

A very important characteristic in the data can explain why the Auto-Regressive and rate of change models were the most accurate. These models all used the relationship between two consecutive quarters in order to determine the predicted quarters. The level assumed that the rate of change between any consecutive quarters is 1. The Quarterly and Seasonal models both assumed the rate of change between the last known quarter and the first predicted quarter was the same as the rate of change between some previous two consecutive quarters. The Auto-Regressive model assumed the rate of change between the last known quarter and the first predicted quarter was a weighted average of the rates of changes between all consecutive quarters in the data sample. Since the most important relationship to determine is between consecutive quarters, it was illogical to try to incorporate the orders beyond the first (like 2nd, 3rd, 4th Orders...).

This aspect of the data can also explain why the other methods ended being not as accurate. For all the Basic models, a trend was mapped between all of the quarters in the data sample. That is, these models never focused on the relationship between just two consecutive quarters. The Linear Exponential model, while still uses the same auto-correlation coefficient as the Auto-Regressive model, used all of the previous quarters to determine the predicted quarters. The other accurate models only used the last quarter to determine the predicted quarters.

6 Conclusion / Recommendation

The most consistent prediction method was the Auto-Regressive model. This model worked best when the auto-correlation coefficient was at least equal to .8—the higher the auto-correlation coefficient the better. Therefore, the data sample which maximized the coefficient would be the best sample to use. However, if no data sample resulted in a high auto-correlation coefficient, then the Level model was the best option.

When analyzing which models were the most accurate, external factors were never taken into account. There were many external factors that could affect the predictions. For example, if there were some reasons to believe a seasonal trend exists, then the Yearly model was an appropriate choice. Likewise, if Hanover believed the trend would sharply increase, then the quarterly method would have been the best choice. The best way to predict future quarters was to use the model which was most applicable with the external factors that existed.

Appendix A: NY_PIP with Data range 1-20

t	x_t
1	1772.67
2	1989.64
3	2466.11
4	2924.42
5	3417.84
6	3699.57
7	3954.49
8	3900.85
9	3904.89
10	4221.17
11	4421.22
12	4729.64
13	5228.22
14	5059.26
15	4897.24
16	4848.63
17	4651.72
18	4503.05
19	4744.14
20	4884.63

Appendix B: Each State's % of total losses

All Coverages	Sum of Losses	% of total	# of claims	% of total	Exposure (# policies)	% of total
CO	\$ 8,820,578.05	0.03%	2,698	0.02%	55,518	0.02%
MD	18,876,452	0.07%	8,388	0.06%	162,375	0.07%
TX	77,344,346	0.28%	31,554	0.24%	671,031	0.28%
RI	92,377,124	0.33%	25,997	0.20%	495,354	0.20%
AR	109,131,746	0.39%	56,045	0.42%	1,276,223	0.53%
WI	137,776,478	0.50%	63,586	0.48%	1,570,327	0.65%
NC	139,759,876	0.50%	59,862	0.45%	1,642,472	0.68%
TN	140,131,450	0.51%	52,118	0.39%	1,229,096	0.51%
OK	147,269,070	0.53%	50,783	0.38%	1,463,083	0.60%
OH	149,084,212	0.54%	61,427	0.46%	1,291,464	0.53%
NH	213,567,461	0.77%	116,777	0.88%	2,133,169	0.88%
VA	246,566,177	0.89%	115,459	0.87%	2,707,879	1.12%
LA	326,376,063	1.18%	133,303	1.00%	2,446,088	1.01%
IL	375,150,117	1.36%	170,518	1.28%	3,607,759	1.49%
GA	376,033,827	1.36%	136,874	1.03%	3,524,676	1.45%
FL	440,224,746	1.59%	149,495	1.12%	4,621,842	1.90%
ME	483,798,841	1.75%	282,512	2.12%	5,695,645	2.35%
IN	589,536,479	2.13%	266,979	2.01%	3,144,698	1.29%
CT	597,252,681	2.16%	228,762	1.72%	4,523,898	1.86%
NY	1,660,466,306	6.00%	573,228	4.31%	13,738,617	5.66%
NJ	2,399,787,853	8.67%	566,288	4.26%	15,806,990	6.51%
MA	6,265,731,453	22.64%	3,140,577	23.60%	43,030,191	17.72%
MI	12,685,660,423	45.83%	7,012,155	52.70%	128,002,991	52.71%
Total	\$ 27,680,723,757.14	100.00%	13,305,385	100.00%	242,841,383	100.00%
MI+MA+NJ+NY:	23,011,646,035	83.13%	11,292,248	84.87%	200,578,789	82.60%

BI	Sum of Losses	% of total	# of claims	% of total	Exposure (# policies)	% of total
CO	\$ 2,008,187.12	0.04%	135	0.03%	12,900	0.03%
MD	4,884,866	0.09%	671	0.13%	35,126	0.08%
TX	21,686,923	0.41%	3,003	0.58%	141,674	0.32%
TN	23,066,688	0.43%	3,143	0.61%	323,075	0.73%
AR	25,183,574	0.47%	2,889	0.56%	314,684	0.71%
RI	34,929,944	0.66%	4,006	0.78%	129,196	0.29%
OK	35,371,960	0.67%	4,437	0.86%	343,113	0.77%
OH	36,255,128	0.68%	3,349	0.65%	285,641	0.64%
WI	37,319,676	0.70%	3,053	0.59%	332,090	0.75%
NC	45,319,908	0.85%	5,897	1.15%	464,294	1.05%
VA	50,140,931	0.94%	6,296	1.22%	641,952	1.44%
NH	63,279,641	1.19%	4,971	0.97%	549,760	1.24%
ME	71,625,123	1.35%	7,165	1.39%	916,881	2.06%
GA	83,155,808	1.56%	11,294	2.19%	848,561	1.91%
IL	94,047,592	1.77%	9,284	1.80%	885,419	1.99%
LA	101,205,540	1.90%	13,193	2.56%	671,663	1.51%
IN	115,845,268	2.18%	11,043	2.14%	712,774	1.60%
FL	122,382,023	2.30%	7,518	1.46%	872,218	1.96%
CT	166,835,473	3.14%	13,527	2.63%	956,175	2.15%
NY	330,836,436	6.22%	12,868	2.50%	2,101,490	4.73%
NJ	625,677,857	11.76%	38,710	7.52%	3,162,526	7.12%
MI	1,101,013,568	20.70%	32,930	6.39%	20,137,780	45.33%
MA	2,127,002,762	39.99%	315,576	61.28%	9,588,398	21.58%
Total	\$ 5,319,074,874.80	100.00%	514,958	100.00%	44,427,390	100.00%
MI+MA+NJ+NY:	4,184,530,624	78.67%	400,084	77.69%	34,990,194	78.76%

CM	Sum of Losses	% of total	# of claims	% of total	Exposure (# policies)	% of total
CO	\$ 937,366.80	0.02%	1,201	0.02%	10,366	0.02%
MD	3,086,660	0.08%	3,069	0.05%	29,683	0.06%
RI	6,085,214	0.15%	5,096	0.08%	109,790	0.23%
TX	9,579,818	0.24%	10,904	0.16%	127,224	0.27%
AR	17,274,919	0.43%	30,739	0.46%	298,937	0.64%
NC	18,356,434	0.46%	26,327	0.40%	367,096	0.78%
TN	18,749,669	0.47%	17,882	0.27%	271,831	0.58%
OK	20,658,920	0.52%	16,037	0.24%	350,441	0.75%
OH	22,125,925	0.56%	26,048	0.39%	323,348	0.69%
WI	22,171,701	0.56%	29,757	0.45%	407,453	0.87%
NH	30,271,578	0.76%	54,372	0.82%	525,125	1.12%
FL	35,453,624	0.89%	48,405	0.73%	881,756	1.88%
VA	37,675,337	0.95%	53,067	0.80%	649,790	1.39%
GA	42,977,843	1.08%	43,704	0.66%	830,382	1.77%
IL	46,876,860	1.18%	72,116	1.09%	886,421	1.89%
LA	55,277,156	1.39%	63,174	0.95%	550,113	1.17%
CT	66,001,417	1.66%	96,799	1.46%	1,148,347	2.45%
ME	78,601,281	1.97%	140,575	2.12%	1,540,561	3.29%
IN	95,443,791	2.40%	116,541	1.76%	791,126	1.69%
NJ	187,290,263	4.70%	153,573	2.32%	2,793,441	5.96%
NY	203,972,928	5.12%	253,108	3.83%	2,790,029	5.96%
MA	742,485,996	18.63%	1,283,619	19.40%	6,850,835	14.62%
MI	2,223,519,376	55.80%	4,070,249	61.52%	24,313,487	51.90%
Total	\$ 3,984,874,076.86	100.00%	6,616,362	100.00%	46,847,578	100.00%
MI+MA+NJ+NY:	3,357,268,563	84.25%	5,760,549	87.07%	36,747,791	78.44%

CO	Sum of Losses	% of total	# of claims	% of total	Exposure (# policies)	% of total
CO	\$ 2,104,815.60	0.02%	629	0.02%	10,323	0.02%
MD	5,454,815	0.06%	2,020	0.05%	29,045	0.06%
RI	20,428,686	0.22%	7,169	0.19%	100,634	0.22%
TX	21,723,272	0.23%	7,073	0.19%	125,340	0.28%
AR	37,777,574	0.41%	11,840	0.32%	291,244	0.65%
WI	40,708,619	0.44%	17,164	0.46%	378,747	0.84%
NC	41,143,446	0.44%	13,594	0.37%	345,787	0.77%
OK	46,216,184	0.50%	14,626	0.39%	346,377	0.77%
OH	46,233,932	0.50%	16,483	0.45%	304,930	0.68%
TN	56,795,802	0.61%	16,466	0.44%	263,165	0.58%
NH	72,073,331	0.78%	32,680	0.88%	487,570	1.08%
VA	81,713,283	0.88%	27,164	0.73%	606,672	1.34%
LA	99,933,201	1.08%	29,500	0.80%	535,267	1.19%
FL	110,557,477	1.19%	41,515	1.12%	859,383	1.90%
GA	119,556,311	1.29%	39,927	1.08%	795,246	1.76%
IL	133,080,804	1.44%	48,901	1.32%	861,755	1.91%
ME	160,673,053	1.74%	71,139	1.92%	1,391,040	3.08%
CT	178,517,032	1.93%	58,620	1.58%	1,030,282	2.28%
IN	202,911,948	2.19%	73,732	1.99%	741,063	1.64%
NY	401,186,986	4.33%	131,084	3.54%	2,533,911	5.62%
NJ	492,795,549	5.32%	147,995	4.00%	2,636,041	5.84%
MA	1,597,933,976	17.26%	685,697	18.52%	7,393,726	16.39%
MI	5,286,211,035	57.11%	2,208,333	59.63%	23,054,181	51.09%
Total	\$ 9,255,731,132.07	100.00%	3,703,351	100.00%	45,121,731	100.00%
MI+MA+NJ+NY:	7,778,127,546	84.04%	3,173,109	85.68%	35,617,859	78.94%

PD	Sum of Losses	% of total	# of claims	% of total	Exposure (# policies)	% of total
CO	\$ 1,302,935.61	0.04%	495	0.03%	12,927	0.03%
MD	3,538,629	0.11%	1,621	0.10%	35,103	0.08%
TX	16,855,811	0.52%	7,513	0.46%	141,609	0.32%
RI	19,184,266	0.60%	8,050	0.49%	129,219	0.29%
WI	20,078,394	0.63%	9,024	0.55%	332,040	0.75%
AR	21,810,256	0.68%	8,683	0.53%	314,746	0.71%
OH	25,801,189	0.80%	11,349	0.69%	285,665	0.64%
OK	28,348,124	0.88%	12,017	0.73%	343,153	0.77%
TN	31,896,664	0.99%	12,784	0.77%	323,061	0.73%
NC	34,940,088	1.09%	14,044	0.85%	464,391	1.04%
NH	44,636,112	1.39%	23,728	1.44%	549,568	1.23%
VA	53,163,952	1.66%	23,086	1.40%	641,947	1.44%
ME	58,647,596	1.83%	29,678	1.80%	916,639	2.06%
LA	66,404,001	2.07%	26,624	1.61%	671,749	1.51%
FL	71,344,235	2.22%	30,266	1.83%	872,675	1.96%
IL	78,769,777	2.45%	35,645	2.16%	885,297	1.99%
GA	82,468,451	2.57%	32,347	1.96%	848,564	1.91%
CT	102,929,377	3.20%	42,978	2.60%	955,338	2.15%
IN	107,672,457	3.35%	48,510	2.94%	712,760	1.60%
NY	197,436,933	6.15%	83,740	5.07%	2,101,376	4.72%
NJ	361,401,791	11.25%	135,434	8.21%	3,162,545	7.10%
MI	414,955,464	12.92%	370,325	22.44%	20,232,613	45.43%
MA	1,368,418,011	42.60%	682,334	41.35%	9,602,924	21.56%
Total	\$ 3,212,004,513.49	100.00%	1,650,275	100.00%	44,535,909	100.00%
MI+MA+NJ+NY:	2,342,212,199	72.92%	1,271,833	77.07%	35,099,458	78.81%

CSL	Sum of Losses	% of total	# of claims	% of total	Exposure (# policies)	% of total
NC		0.00%		0.00%	905	0.01%
CO	182,969	0.01%	38	0.02%	676	0.01%
NH	3,306,799	0.23%	1,026	0.44%	21,146	0.20%
LA	3,556,164	0.25%	812	0.35%	17,296	0.16%
TX	3,562,715	0.25%	1,047	0.45%	16,792	0.16%
AR	7,085,423	0.49%	1,894	0.81%	56,612	0.54%
TN	9,622,627	0.67%	1,843	0.78%	47,965	0.46%
RI	11,749,015	0.82%	1,676	0.71%	26,516	0.25%
OK	16,673,882	1.16%	3,666	1.56%	79,999	0.76%
WI	17,498,088	1.22%	4,588	1.95%	119,996	1.14%
OH	18,668,037	1.30%	4,198	1.79%	91,880	0.88%
IL	22,375,084	1.56%	4,572	1.95%	88,866	0.85%
VA	23,872,675	1.66%	5,846	2.49%	167,518	1.60%
FL	24,192,738	1.69%	4,941	2.10%	133,043	1.27%
WI	40,708,619	2.84%	17,164	7.31%	378,747	3.61%
GA	47,875,414	3.34%	9,602	4.09%	201,923	1.93%
IN	67,663,016	4.72%	17,153	7.31%	186,975	1.78%
CT	77,995,520	5.44%	15,122	6.44%	320,733	3.06%
VA	81,713,283	5.70%	27,164	11.57%	606,672	5.78%
NJ	112,615,720	7.85%	17,254	7.35%	444,741	4.24%
ME	114,251,788	7.97%	33,955	14.46%	930,523	8.87%
NY	271,559,088	18.94%	42,339	18.03%	1,054,888	10.06%
MI	457,171,641	31.88%	18,878	8.04%	5,493,990	52.38%
Total	\$ 1,433,900,302.96	100.00%	234,778	100.00%	10,488,402	100.00%
MI+ME+NJ+NY:	955,598,236	66.64%	112,426	47.89%	7,924,142	75.55%

PIP	Sum of Losses	% of total	# of claims	% of total	Exposure (# policies)	% of total
MD	\$ 1,911,481.87	0.04%	1,007	0.16%	33,418	0.06%
CO	2,284,304	0.05%	200	0.03%	8,326	0.02%
TX	3,935,807	0.09%	2,014	0.32%	118,391	0.23%
CT	4,973,862	0.11%	1,716	0.27%	113,022	0.22%
FL	76,294,649	1.66%	16,850	2.67%	1,002,767	1.91%
NY	255,473,936	5.56%	50,089	7.95%	3,156,923	6.02%
MA	429,890,707	9.35%	173,351	27.52%	9,594,308	18.31%
NJ	620,006,673	13.49%	73,322	11.64%	3,607,696	6.88%
MI	3,202,789,339	69.66%	311,440	49.44%	34,770,941	66.35%
Total	\$ 4,597,560,758.80	100.00%	629,989	100.00%	52,405,792	100.00%
MI+MA+NJ+NY:	3,202,789,339	98.06%	311,440	96.54%	34,770,941	97.57%

Appendix C: Detailed Model Test

Representation of each number:

1	2	3	4	5	6	7	8	9
Linear	Exponential	Log	Power	Yearly	Level	Quarterly	Auto- Regressive	Linear Exponential

MA_Frequency										MA_Severity									MA_Pure Premium								
Data Range	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
5 to 34								1							1									1			
20 to 44						1											1							1			
30 to 44						1											1							1			
25 to 44						1									1									1			
15 to 44						1											1							1			
10 to 39						1								1										1			
20 to 39							1							1									1				
Total	0	0	0	0	0	5	1	1	0	0	0	0	0	2	2	0	3	0	0	0	0	0	1	6	0	0	0

MI_Frequency										MI_Severity									MI_Pure Premium								
Data Range	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
5 to 34							1									1									1		
20 to 44								1							1					1							
30 to 44								1		1																1	
25 to 44								1									1									1	
15 to 44								1							1											1	
10 to 39							1								1									1			
20 to 39							1				1								1								
Total	0	0	0	0	0	0	3	4	0	1	1	0	0	0	3	1	1	0	1	0	1	0	0	1	1	3	0

NY_Frequency										NY_Severity										NY_Pure Premium									
Data Range	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9		
5 to 34						1				1																	1		
20 to 44								1							1									1					
30 to 44							1								1									1					
25 to 44							1								1									1					
15 to 44								1							1									1					
10 to 39								1									1									1			
20 to 39						1									1											1			
Total	0	0	0	0	0	2	2	3	0	0	1	0	0	0	5	0	1	0	0	0	0	0	0	4	0	3	0		

NJ_Frequency										NJ_Severity										NJ_Pure Premium									
Data Range	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9		
5 to 34								1							1									1					
20 to 44						1											1										1		
30 to 44						1				1																1			
25 to 44						1				1																	1		
15 to 44						1					1									1									
10 to 39						1											1							1					
20 to 39						1											1							1					
Total	0	0	0	0	0	6	0	1	0	2	1	0	0	0	1	0	3	0	0	1	0	0	0	3	0	1	2		

Reference

1. Our promise to you is world-class performance when and where you need it. The Hanover Insurance Group, Inc. 2009. <http://www.hanover.com/thg/about/index.htm> (accessed April 27, 2009).
2. *exposure base*. International Risk Management Institute, Inc. 2000-2009. <http://www.irmi.com/online/insurance-glossary/terms/e/exposure-base.aspx> (accessed April 2, 2009).
3. *frequency*. International Risk Management Institute, Inc. 2000-2009. <http://www.irmi.com/online/insurance-glossary/terms/f/frequency.aspx> (accessed April 2, 2009).
4. *Insurance Coverage Definitions | Car Insurance Coverage*. 2009. <http://www.carinsurance.com/CoverageDefinitions.aspx>. (accessed April 2, 2009).
5. *Insurance Coverage Definitions | Car Insurance Coverage*. 2009. <http://www.carinsurance.com/CoverageDefinitions.aspx> (accessed April 2, 2009).
6. *loss costs*. International Risk Management Institute, Inc. 2000-2009. <http://www.irmi.com/online/insurance-glossary/terms/l/loss-costs.aspx> (accessed April 2, 2009).
7. In *Forecasting With Univariate Box-Jenkins Models*, by Alan Pankratz. New York: John Wiley & Sons, Inc, 1983.
8. *What is CSL coverage, required to lease a car in California?* 2009. <http://www.carinsurance.com/kb/content37475.aspx> (accessed April 4, 2009).
9. William M.K. Trochim. *General Linear Model*. October 20, 2006. <http://www.socialresearchmethods.net/kb/genlin.php> (accessed April 4, 2009).