SOCIETY OF ACTUARIES

Predictive Modeling: A Modeler's Introspection

A Paper Describing How to Model and How to Think Like a Modeler

June 2015

SOCIETY OF ACTUARIES

Predictive Modeling: A Modeler's Introspection

A Paper Describing How to Model and How to Think Like a Modeler

SPONSORS Committee on Finance Research

AUTHORS Michael Ewald, FSA, CFA, CERA
Qiao Wang

Acknowledgements

The authors would like to express their gratitude to the SOA Project Oversight Group. The quality of this paper has been enhanced through their valuable input. The group includes the following individuals: William Cember, Andy Ferris, Jean-Marc Fix (chair), John Hegstrom, Christine Hofbeck, Steve Marco, Dennis Radliff, Barbara Scott (SOA), and Steven Siegel (SOA).

Additionally, we would like to thank our colleagues and friends: Andrew Dalton, Katherine Ewald, Jim King, Alex Marek, and Tom Tipton. We could not have produced this paper without their help and feedback. Additionally, they have helped make us better modelers, actuaries, and people.

Table of Contents

Introduction – Historical Actuarial Approach

Insurance practitioners have been analyzing risk for thousands of years. From the Code of Hammurabi, which waived loans on a ship if it was lost in voyage, to the initial mortality tables developed by John Graunt in the mid-18th century[1], the goal of the actuary has remained constant: analyze "the financial consequences of risk."[2] As society has progressed and technology has developed, both the design of insurance products and the means of analyzing the risk have increased in complexity. The introduction of computers has, in a very short period of time, changed the way actuaries approach their day-to-day jobs. The need for both technical and product expertise has never been greater. The goal of this paper is to help you understand one tool that has gained an enormous amount of traction over the past two decades: predictive modeling. Predictive modeling is the practice of leveraging statistics to predict outcomes. The topic covers everything from simple linear regression to machine learning. The focus of this paper is a branch of predictive modeling that has proven extremely practical in the context of insurance: Generalized Linear Models (GLMs).

Before moving to GLMs, it is important to understand actuarial techniques that have been used in the past. In particular, we will discuss one and two-way analysis. Although the terminology may be foreign, anyone with a basic analytical background has used these techniques. One-way analysis refers to the review of a response by a single variable (e.g. observed mortality by age). Two-way analysis looks at the response by two variables (e.g. observed mortality by age and gender). Relativities between variable groupings are then used to ascertain the risk between the two groups (i.e. mortality of a 46 year old male versus that of a 45 year old male). This technique is then applied to other variables to help segment the risk in question. Although this technique is extremely intuitive, it has a number of well-known drawbacks:

- Ignores correlations - If Detroit experience is 20% worse than average and if auto manufacturing segment experience is 20% worse than average, should we expect auto manufacturers in Detroit to have experience that is 40% worse than average?
- Suffers from sequencing bias - The first variable analyzed may account for the signal of variables that will be analyzed later. This could reduce or eliminate the importance of those variables.
- Cannot systematically identify noise – One large claim or a large unique policy can have a very large impact on one-way analyses. It is difficult to identify the signal (true impact of a variable) vs. noise (volatility) in one-way analyses.

[1] Klugman, Stuart A., "Understanding Actuarial Practice," Society of Actuaries, 2012, Page 7.
[2] Society of Actuaries. 2010. "What is an Actuary?" https://www.soa.org/about/about-what-is-an-actuary.aspx.

Although actuaries have developed techniques to account for these pitfalls, they are time consuming and require substantial judgment. GLMs provide a systematic approach that addresses the above concerns by allowing all variables to be analyzed in concert.

GLM Overview

Let us begin with normal linear regression. Most readers will be familiar with this form of GLM. It is well known that linear models, along with all GLMs, require independence of observations. There are three assumptions unique to linear regression that we address below[3]:

1. Error terms follow normal distribution
2. Variance is constant
3. The covariates effect on the response are additive

These assumptions are relaxed when reviewing GLMs. The new assumptions are as follows[4]:

1. Error term can follow a number of different distributions from the exponential family
2. Variance does not need to be constant
3. The covariates can be transformed so that their effect is not required to be additive.

Relaxing the first assumption allows us to utilize exponential distributions that apply directly to the insurance industry. For distributions that are strictly non-negative, such as claims, a normal distribution that exists across all real numbers is not ideal. The Poisson distribution and Gamma distribution apply to only positive numbers and are more appropriate for claim counts and claim amounts, respectively. The Binomial distribution applies to all binary datasets, and may be appropriate for modeling policyholder behavior or mortality.

GLMs allow variance to adjust with the mean. As you can see in the exhibit below, the variance of the Poisson and Gamma distributions increase as their means increase:

	Mean	Variance
Poisson	θ	θ
Gamma	$\alpha\theta$	$\alpha\theta^2$
Binomial	np	np(1-p)
Normal	μ	σ^2

There are many instances where this makes intuitive sense. When modeling claim amounts, we expect absolute variability to increase as the dollars increase. Additionally, when reviewing the binomial variance formula, variance approaches zero as the mean tends toward zero or one.[5]

[3] McCullagh, P. and J. A. Nelder, "Generalized Linear Models", 2nd Edition, Chapman & Hall, CRC, 1989. Pages 23.
[4] McCullagh, P. and J. A. Nelder, "Generalized Linear Models", 2nd Edition, Chapman & Hall, CRC, 1989. Pages 31.
[5] McCullagh, P. and J. A. Nelder, "Generalized Linear Models", 2nd Edition, Chapman & Hall, CRC, 1989. Pages 12.

Finally, the assumption of additive covariates is neither intuitive, nor can it be easily applied to actuarial analyses. Predictive modeling allows this assumption to be relaxed. For example, risk factors can be multiplicative. This is extremely useful when developing a pricing algorithm that follows a manual rate structure.

The Math

Before addressing the calculations to develop a model, it is important to understand the structure of the final equation. There are three main components of both normal linear regression and GLMs:

1. Random Component – distribution of the error term.
2. Systematic Component – summation of all covariates (i.e. predictors) that develop the predicted value.
3. Link Function – relates the linear predictor to the expected value of the dataset. In other words, it is a mathematical transformation of the linear predictor that allows the expected mean to be calculated.

The linear regression model is as follows:

$$E[Y_i] = \underbrace{\beta_0 + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \cdots + \beta_n * x_{i,n}}_{\text{Systematic Component}} + \underbrace{\varepsilon_i}_{\text{Random Component}}$$

The formula above shows the expected value of each observation i. This is calculated by multiplying a vector of rating variables B and a vector of observed values X. The notation will make more sense in later sections when these formulas are applied to real life scenarios. In linear regression, the expected value is equal to the systematic component because the identity link (described below) does not transform the systematic component.

When moving to a GLM, the random component can take on a member of the exponential family and the systematic component is transformed via a link function:

$$E[Y_i] = \underbrace{g^{-1}}_{\text{Link Function}}(\underbrace{\beta_0 + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \cdots + \beta_n * x_{i,n}}_{\text{Systematic Component}}) + \underbrace{\varepsilon_i}_{\text{Random Component}}$$

The link function is simply an algebraic transformation of the systematic component. Please see the applications section below for examples of the link function.

Now that we have seen what we are working towards, let us discuss how we develop the final formula. GLMs develop parameters in such a way that minimizes the difference between actual and predicted values. In linear regression, parameters are developed by minimizing the sum of squared errors. The same result can be achieved through maximum likelihood estimation (MLE). MLEs are used to parameterize GLMs.

For any member of the exponential family there exists a density function: $f(y;\theta)$. Taking the log of each density function for each observation and adding them together gives the log likelihood function.[6]

$$l(\mu;y) = \sum_i \log f_i(y_i;\theta_i)$$

To calculate the parameters, take the derivative of the likelihood function with respect to the rating variables and set them to zero:

$$\frac{\partial l}{\partial \beta_1} = 0$$

$$\frac{\partial l}{\partial \beta_2} = 0$$

$$\vdots$$

$$\frac{\partial l}{\partial \beta_n} = 0$$

Using these equations, you then solve for the parameters in question.

Applications

Distributions

The type of GLM chosen for a project hinges on the distribution of the data. The table below details the major distributions.[7]

The table below details the canonical link for some common distributions. The canonical link is the link that can be used to calculate the parameters in closed form. Given the size of modeling

[6] McCullagh, P. and J. A. Nelder, "Generalized Linear Models", 2nd Edition, Chapman & Hall, CRC, 1989. Page 24.
[7] McCullagh, P. and J. A. Nelder, "Generalized Linear Models", 2nd Edition, Chapman & Hall, CRC, 1989. Page 30.

datasets, parameters are never calculated in this manner. They are calculated using numerical methods (i.e. an iterative process). Since the canonical link is not a requirement to parameterize the model, other link functions may be chosen if they are more practical.

	Range of y	Canonical Link	E(Y;θ)	Var Function
Normal	$(-\infty,\infty)$	identity	θ	1
Poisson	0, 1, … ∞	log	$\exp(\theta)$	μ
Binomial	0, 1	logit	$e^{\theta}/(1+e^{\theta})$	$\mu(1-\mu)$
Gamma	$(0,\infty)$	reciprocal*	$1/\theta$	μ^2

* Gamma distribution often uses log link function for easier implementation. This is explained in more detail in subsequent sections.

The name of the link functions and their transformations are described in the "Canonical Link" and "E(Y;θ)" columns. For example, the binomial distribution uses the logit link. The transformation is as follows:

$$\frac{e^{Systematic\ Component}}{1 + e^{Systematic\ Component}}$$

Normal: The normal distribution is useful for modeling any response that exists between negative and positive infinity. It can be used in an analysis focused on changes in a variable. For example, one might want to look at the drivers of changes in premium. This practice is also known as premium dislocation.

Poisson: The Poisson distribution is primarily used to model claim counts when multiple claims can occur in a given exposure period. For example, a policyholder with health care coverage can have multiple physician claims in a given month. The Poisson distribution is also useful when modeling claims in group insurance, because claim counts may be reviewed at the group level rather than the individual level.

Binomial: The binomial distribution can be used to model binary responses. Examples in an insurance context are as follows:

- Mortality – the probability of dying.
- Lapse/persistency – probability of a policyholder/client retaining or cancelling their coverage.
- Close ratio/issue rate – probability of selling a policy.
- Annuitization – probability of a policyholder annuitizing a contract.

- Termination/recovery – probability a claimant currently on disability will recover and the claim will terminate. This can be used to develop reserve assumptions.

The binomial distribution is by no means limited to the examples above. The distribution can be used for any other situation characterized by a binary response. Moreover, non-binary data can be transformed into a binary distribution to develop meaningful analytics. For example, when reviewing large claim amounts, you can create a binary distribution that looks at claims greater than a certain threshold to determine the characteristics of those claims.

Gamma: The gamma distribution is most often used for claim severity (i.e. claim amount given a claim has occurred). The gamma distribution is appropriate to handle the long tailed nature of claim amounts. Additionally, it can be used to model pure premium (loss costs excluding overhead and profit loading).

Tweedie: Although not shown in the table above, the Tweedie distribution is very useful in both P&C and Group insurance. The Tweedie distribution is a hybrid between the Poisson and Gamma distributions. Therefore, it is applicable to pure premium and loss ratio modeling. Loss ratio modeling determines what variables cause deviation from the average loss ratio. For cases with minimal exposures, there is high probability of zero claims and a loss ratio of zero. The Tweedie distribution, with a point-mass at zero and a long tail, is perfect for this type of analysis.

We detail the common distributions and their potential uses below:

	Potential Models	Link Function
Normal	Dislocation (i.e. Change in Premium)	identity
Poisson	Claim Frequency (incidence)	log
Binomial	Mortality Rates Lapse/Persistency Rates Close Ratio/Issue Rates Annuitization Rates Termination/Recovery Rates	logit
Gamma	Claim Severity Pure Premium	log*
Tweedie	Pure Premium Loss Ratio Modeling	log

* Gamma distribution often uses log link function rather than the reciprocal link. This is explained in more detail in subsequent sections.

Good Models

Choosing the appropriate distribution is paramount when embarking on a modeling project. However, choosing an appropriate distribution does not necessarily lead to a good model. The modeler will utilize historical information to inform decisions regarding future events. A good

model will reflect historical information expected to persist in the future and control for historical anomalies. This involves avoiding two opposing modeling risks: overfitting and underfitting.

Overfitting occurs when a model does a good job of predicting the past, but a poor job of predicting the future. An extreme example of overfitting is creating a model with as many parameters as observations. The model would predict the past perfectly, but it would not provide much insight to future.[8]

Underfitting does a good job of predicting the future on average, but does not do a good job of segmenting risk factors. An extreme example is a model with one parameter: the average of the past.

The modeling process discussed in subsequent sections will help the reader avoid both overfitting and underfitting. Other principles mentioned in the Generalized Linear Model textbook by P. McCullagh and J. A. Nelder are as follows[9]:

1. "… all models are wrong; some, though, are more useful than others and we should seek those."
2. Do not "… fall in love with one model to the exclusion of alternatives."
3. The modeler should perform "… thorough checks on the fit of a model."

The first principle lays the groundwork for a skepticism that all modelers should possess. The popularity of modeling, driven by the results it has produced over the past decade, leads most individuals to give significant (and possibly too much) credence to the practice. Although it is a very powerful tool, modeling is only as good as the modeler's ability to take past data and inform meaningful future results.

The second point explains that modeling may not always be the best solution. Time constraints, lack of data, lack of resources, etc. are all impediments to developing a good model. Do not blindly choose a predictive model without understanding the costs and benefits.

The final point will be discussed further in the "Model Validation" section.

Data Overview

The first, and arguably the most important aspect of predictive modeling is data preparation. We often call this "the 80%" because 80% of the time will be spent preparing the data and 20% will be spent building and checking the model. This is not a strict rule. Given that the majority of the authors' work has been the first of its kind, data preparation is very time consuming. As companies focus more and more on "model ready data," the goal is that data preparation will

[8] McCullagh, P. and J. A. Nelder. "Generalized Linear Models." 2nd Edition. Chapman & Hall, CRC. 1989. Page 7.
[9] McCullagh, P. and J. A. Nelder, "Generalized Linear Models", 2nd Edition, Chapman & Hall, CRC, 1989. Page 8.

take 20% of the time and modeling will take 80%, with the absolute amount of time per project decreasing substantially.

The modeling dataset consists of rows, which are observations. Each observation will have a corresponding exposure or exposure unit (i.e. weight). Depending on the dataset, observations may equal exposures. For example, an individual life mortality study may have a row for each life month. In this case, the number of observations equals the number of exposures. In other instances, exposures will differ from observations. In a loss ratio model, which compares claim amounts to premium for clients over a period of time, the exposure would be the premium paid in the period that corresponds to the claim amounts.

The dataset columns consist of three sections: response, weight, and variables. The response field is also referred to as the dependent variable. The dependent variable is what you are trying to model, and the independent variables are what you are using to predict the dependent variable. The weight defines the exposure (i.e. credibility) of each observation. The more weight of an observation means it will have a larger impact on the final parameters than an observation with less weight. The independent variables (variables 1, 2, 3, etc.) in the exhibit below are what will be tested for their predictive power; they are often referred to as covariates.

	Response	Weight	Variable 1	Variable 2	Variable 3	…	Variable n
Observation 1							
Observation 2							
Observation 3							
⋮							
Observation n							

Modeling Process

Project Scope

Laying out the initial project scope is imperative for any modeling project. This is most important when embarking on a predictive model for the first time. Interested in your new work, your business partners will probably ask for additional analysis. Sticking to a pre-defined scope will limit project creep and allow you to meet your deliverables in a timely manner.

Modeling Participants:

Modeling Team	- Modeling experts - Need strong data and business knowledge
Business Partners	- Have business knowledge (product, claims, underwriting, sales, etc.) of data and business process - Owners of the final model

	- Their involvement in the process helps them better understand and support the final model - Advocate your model across organization
IT	- Produce initial dataset - Implement final model
Project Manager	- Provide management expertise, allowing other participants to spend time more efficiently

While setting scope, it is important to define the relationships with project participants, especially IT. When dealing with large amounts of unstructured data, IT will be an important resource for both acquiring and understanding the data. All too often, data is assumed to be clean. However, the modeler should understand that data is often entered by individuals whose goals are not necessarily aligned with the most accurate model data. Underwriters, sales representatives, and clients are often the major suppliers of data.

Given outdated IT systems, constraints, and other impediments, these business partners may enter information that results in the correct answer in their context, but that does not lead to accurate data. Group life insurance provides a good example. Benefit amounts can be defined as either flat or a multiple of salary. If a policyholder chooses a flat benefit and salary is not rate bearing, the underwriter may enter the flat benefit amount as the salary. Although this has no price implications, it would make it difficult for the modeler to understand the true impact of salary on mortality. In group insurance, we check the percent of people with the same "salary." If this number is large, the underwriter most likely entered a flat benefit amount for each individual. This data should be segregated from the true salary data. It may be useful to include sales and underwriting in portions of this process to better understand the data.

Along with the benefit of understanding data, including representatives from sales, underwriting, and other users of your model may have other benefits. If sales will be impacted, it is important to include them in the process so that they are aware of what is changing and why. If they feel included, they will be advocates for you and will help socialize the changes with the field. This must be delicately balanced with the risk of having too many people involved. The modeling team and business partners must decide the best means of balancing the pros and cons of including groups affected by the project. Depending on the scope of the project, other business partners may include, but are not limited to, claims, ERM, marketing, and underwriting staff.

Data Collection

Data Scope

The first step when compiling a dataset is to determine the scope of variables that will be tested in the model. Naturally, you will want to test internal data sources. Additionally, you may want to leverage external data to supplement the internal data. Common sources include government

data, credit data, social networking websites, Google, etc. This data is commonly acquired through third party data aggregators, including Experian, Axiom, D&B, and others. External data can also be used to inform data groupings or as a proxy for other variables, such as a geographic proxy. Additionally, external data can be combined with internal data to create interesting variables. For example, you could combine salary and median income in a region to control for cost of living across regions. A $100,000 salary in New York City leads to a very different lifestyle than $100,000 in North Dakota.

It is important to avoid preconceived notions when constructing the dataset. Including as much information (where practical) in the beginning will allow for the most meaningful analysis. More variables generally lead to more "a-ha" moments. Additionally, there is very little harm including additional data fields, because the modeling process will eliminate insignificant variables.

Data Structure

After outlining the data scope, it is important to define the structure of the data. You need to define the exposure period of the response. For example, are you reviewing a policyholder over a year, month, or some other measurement of time? There is no pre-defined answer, but your decision will depend on a number of factors:

- Amount of data – if you have a very large dataset, using monthly data may cause run time issues while transforming variables, joining datasets, and performing analysis. Additionally, the software that you are using to parameterize your models may not be able to handle that much data. The data phase will have multiple iterations as variables are added, questions are answered, and issues are resolved. An extremely large dataset could be very time consuming.
- Seasonality – if you believe the response varies by month, it may be better to take a more detailed view. For example, policyholders are less likely to file a disability claim in the fourth quarter. In pricing, when contracts are guaranteed in yearly increments, seasonality may not make a difference. However, seasonality may be useful when analyzing cash flows.

Data Scrubbing

Once you acquire the data, it is important to determine reasonability. The following analyses should be performed when receiving a new set of data:

- Check the distribution of the response, weight, and potential modeling variables. Review histograms of continuous variables and check the frequency of all discrete variables. It is also helpful to review the response across each variable. Reviewing information with business partners will help ensure that the data is in line with expectations.
- Outlier analysis – review the largest and smallest variables for the response and all numeric variables. Reviewing these outliers can point to observations that you want to

exclude from your analysis. Additionally, you may want to floor or cap variables to more realistic values. For example, if exposures are reinsured above a certain amount, you may want to cap values so they only represent retained exposure. It is important to note that sometimes these outliers should be included in the analysis because you may want to determine why these outliers occurred.

- Date reasonability – a good way to gain comfort in your data is to see if events occur within the expected exposure period. For example, a claim that occurred prior to the policy effective date would set off red flags.
- Null values – look at instances where levels of the data are null. Often times, this is a sign that those records may have integrity issues. You may also be able to find chronic issues with data processing that can be addressed.
- Unreliable data – some data will fail to pass the smell test and should be separated from the variable in question (See same salary example above).
- Default values – IT systems may default values to 999999 or some other variation. These values should be separated into a missing or unknown bucket for a particular variable.

Identifying errors will help reduce the likelihood of spurious results. Although some of these checks may seem trivial, you will be amazed at what you find when digging into the data. In almost all modeling projects, there will be some finding in the data phase that spawns additional analysis, process improvements, or uncovers a previously unknown issue. When scoping out your project, it is imperative that you have sufficient time to dig through the data.

Data Preparation

Once you have sufficiently reviewed the data, it is time to prepare your data for modeling.

Variable Transformations

Transformations are a mathematical formula that adjusts the variable under review. Some common transformations are:

- Variable change over time - for example, change in premium from one point in time to another is extremely useful when looking at client retention.
- Difference between two variables - for example, reviewing individual salary versus median salary is a useful metric to proxy an individual's lifestyle.
- Percentages – for example, reviewing the percent of females or males in a company.
- Duration – calculating the difference between two dates.
- Reciprocal – you may want to reverse the variable for explanatory purposes (e.g. BLS gives output/hour but it may be easier to explain hours per unit of output).[10]
- Logarithm or square root – used to limit the skew of a variable.[11]

[10] Cox, Nicholas J. 2005. "Transformations: An Introduction." http://fmwww.bc.edu/repec/bocode/t/transint.html.
[11] Cox, Nicholas J. 2005. "Transformations: An Introduction." http://fmwww.bc.edu/repec/bocode/t/transint.html.

Variable Grouping

When grouping variables in the data phase, it is important to consider the following:

- Work with your business partners to develop variable groupings that can eventually be implemented. For example, you may use age groupings in line with your current rating structure for ease of implementation. Surrender charge periods in an annuity lapse model would need to be calculated by subtracting effective date from date of exposure.
- Low exposures – if you have levels that relate to a policy provision that is very uncommon, you may not have enough exposure to derive a meaningful result. In these instances, it may be useful to group all of these odd provisions into a single category. If it is determined that these odd provisions impact the response in a manner that could adversely impact the signal of other variables, it would be prudent to remove these observations from the data.
- Create multiple groupings – when reviewing something with many levels such as an individual's age, it may be useful to create additional variables for the purposes of summary analyses. These grouped variables may also be useful when testing interactions, which will be discussed in subsequent sections.
- Group by exposure size – when reviewing a variable like industry classification codes (SIC or NAIC) where the 4-digit code is a more granular version of the 3-digit, 2-digit, and 1 digit codes, you may want to review the 20 largest 3-digit SIC codes and group all other codes together. There are an unlimited number of potential groupings, so the modeler must use their knowledge of the data to inform groupings.
- Group by percentiles – percentiles of continuous variables allow the modeler a systematic means of giving equal credibility to each bucket.

The initial groupings are by no means final. In fact, you will most likely adjust your variable groupings multiple times throughout the model building process.

Separating Datasets

Overview

Predictive modeling is such a powerful tool because there is a systematic approach for validating your algorithms. Two techniques that exist are out-of-sample testing and out-of-time testing.

Out-of-Sample Testing

Out-of-sample testing separates the data into two independent sets. We define the two datasets as train and test. Train is the data that we will use to build our model. Test, as the name implies, is the data that we will use to validate our model. Separating the data into these two datasets can be accomplished through random sampling, stratified sampling, or bootstrapping.

- Random sampling – random sampling can be performed using a random number generator for each observation. If you want 50% of your data in train and 50% in test,

every randomly generated value less or equal to 0.5 could be placed in train and the rest could be placed in the test dataset.

- Stratified sampling – stratified sampling is similar to random sampling. The biggest difference is that the data is split into N distinct groups called stratas. It is up to the modeler to define the stratas. These will often be defined by a discrete variable in the dataset (e.g. industry, case size, region, etc.). Observations from each strata will then be chosen to generate the train dataset. For example, 100,000 observations could be split into 3 stratas:
    - Strata 1 – 50,000 observations
    - Strata 2 – 30,000 observations
    - Strata 3 – 20,000 observations

  You would then take random samples from each strata so that you have 50% of your train and test data from strata 1, 30% from strata 2, and 20% from strata 3.
- Bootstrapping – the biggest difference between bootstrapping and the other two sampling techniques is that bootstrapping allows an observation to exist in multiple datasets. Bootstrapping is rarely used because the main goal of validation is testing your model on an independent dataset. Including the same observation in multiple datasets does not allow for independence. The main use of bootstrapping is to develop many datasets and have a range of parameter values to determine variability and a confidence interval for those parameters.

Regardless of the approach, it is important to make sure that the datasets are indeed independent. If your data is structured in a way that has observations over time (e.g. policy/months) you will have multiple rows for a particular policy. If you proceed with a random sampling of each observation (row) you will have the same policy in both your test and train data. The danger is that you increase the chances of overfitting your model (i.e. validating a result that may not be true in the future). This is especially dangerous if you have a very large unique policy.

Out-of-Time Testing

Out-of-time testing is a very powerful form of validation as well as a good method to market the power of your model. If your dataset is large enough, it is beneficial to hold out a year of data to determine if the model built on other years does a good job of segmenting risk in the year that was held out from model building. If the model performs well, it will be easier to convince your business partners that the model should be implemented. Conversely, if the model does not perform well, it gives an indication that the model requires additional work or that alternative methods should be considered.

Another version of train/test that could also be used to test your models is separating the data by year. In this method, you separate your dataset into two distinct time periods. For example, if your data spans exposure years 2008-2013, you could build a model on 2008-2010 data and test

it on 2011-2013 data. This allows any biases that occurred during certain time periods to be identified.

A thorough understanding of the data and the business is important when determining the year stratification. For example, building and testing a model on data during a recession may not be indicative of the post-recession world. In this instance, you could build the model on "recession" data and test it on "post-recession" data. Even if you don't have a substantial amount of data for one time period, it is still important to perform these tests. If certain factors are not validated, you will use business judgment to determine if they are not validated because of light data (i.e. volatility) or because there is a bias in these time periods.

Dataset Size

After determining the type of sampling, you will need to determine the percent of data to include in the train and test data. As a rule of thumb, the largest datasets will result in a 50/50 split. The smaller the dataset, the more you move towards a 70% train and 30% test dataset. The rationale is that when testing a small dataset, you need to have sufficient observations in your training dataset to develop a model. When reviewing the model on the test dataset, you will expect significant volatility in the actual values, but the actuals should still be centered around the average predicted values.

Dataset size is not the only determining factor. Response volatility will dictate your choice. The split of train and test depends on the circumstances of the project and will vary from model to model.

Finally, if you have an excessive amount of data, you can create more than two datasets. Multiple test datasets will increase your level of comfort with the final model.

Building a Model

For the purposes of illustrating the model building process, we will review a long-term disability insurance pricing project. This includes both claim counts and their respective claim amounts. We chose this example because it details the process of modeling both incidence and severity. This modeling exercise explains nuances not present when modeling mortality or lapse rates. The actual response is not entirely important, because this process can be applied to any analysis across many disciplines.

The modeling dataset consists of experience from 2007 – 2013. We chose to hold out the 2013 experience for out-of-time sampling. The 2007 – 2012 was split so that 60% of the data would be used to train the model and 40% would be used to test the model.

Defining the Distribution

Once you have constructed your initial dataset, it is time to determine the distribution of your data. As mentioned above, insurance data often follows well known distributions. If you are not

modeling a previously identified distribution, review the histograms of your response. This will help you understand what link and error structure to use.

Once you have defined your distribution, use the tables in the Distributions section to determine the link of your data. In most instances, the canonical link is the link that should be chosen for a specific error structure. However, as we discussed, the canonical link may not be practical. For example, we chose the log link rather than the canonical link (reciprocal) for the gamma distribution. Combining the incidence and severity models is easier when both use the log link function because the factors and intercept are multiplicative. This will become clearer in subsequent sections.

Fitting Main Effects

The dataset has been built and the error and link have been defined. It is finally time to start building your model by testing the main effects. Main effects are the variables that will help predict the response in question. For example, age of insured is a main effect.

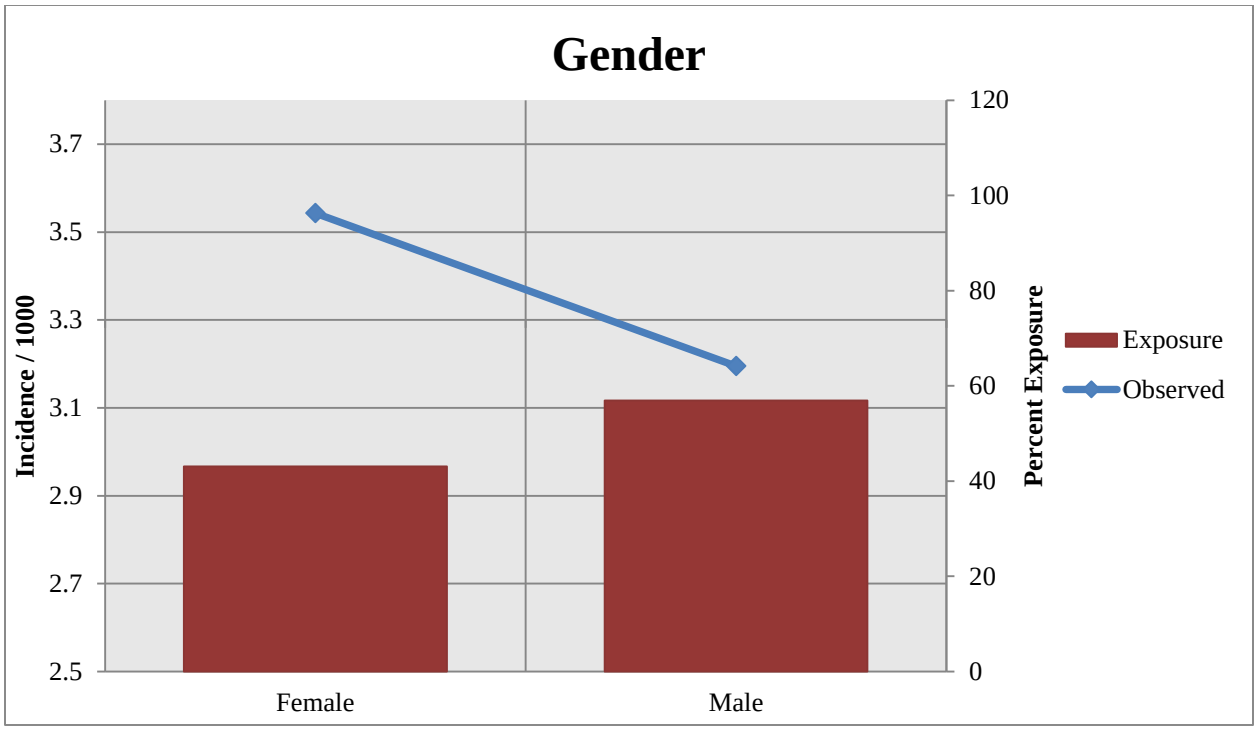
Your review of the data and your business knowledge should give you a strong indication of the most important variables. Fitting these parameters first will reduce the likelihood of adding variables that will be removed later.

The first step in model building is to test the predictive power of each variable. This can be accomplished by fitting simple factors. Simple factors are parameters for each level of a variable. The software you use will dictate the mechanics of adding variables. The examples below will further explain simple factors.

The first indication of the predictive power of your model is the change in deviance. When adding a variable, you would expect a deviance reduction (i.e. the actual and predicted values move closer to one another). A statistical measure that utilizes deviance reduction is AIC (Akaike information criterion). AIC is a deviance measure that penalizes for the number of parameters. If this metric is positive, it implies that the deviance reduction (if there is a deviance reduction) was not enough to justify the increase in parameters.

The Chi-square statistic is a well-known statistical measure that measures the significance of a variable. Generally, a threshold of 5% is used. If you add a variable and the resulting Chi-square statistic is 5%, we have 95% confidence that that the model is statistically different from the original model. Anything above 5% is deemed insignificant and anything below 5% is deemed significant. Depending on the response in question, these thresholds can be increased or decreased.

Now that we have discussed the criteria for adding a variable, let us see it in practice. The first main effect we review is gender. As depicted in the graph below, female incidence is 11% higher than male incidence (3.54 vs. 3.20).



The table below shows the model before and after adding the gender variable:

	Intercept Only Model	Model w/ Gender	Change
DoF	1,402,896	1,402,895	(1)
Parameters	1	2	1
Deviance	642,027,500	641,878,500	(149,000)
AIC	(23,374,820)	(23,523,740)	(148,920)
Chi-square			0.00%

The table above details two models: the intercept only model and the model with Gender. Before explaining the results, we need to define the term intercept. Intercept is the variable in the model that allows the average of your predicted values to equal the average of the actual values. As you add variables to the model, the intercept will move up and down to make sure actuals equal predicted values in aggregate.

Adding the variable gender reduces degrees of freedom (DoF) and increases the number of parameters by 1. This is because only 1 level (in this case female) will have a parameter that distinguishes it from the base. Since no other variables have been added to the model, the predicted value of Males will be based solely on the intercept value.

The Male grouping is defined as the base because modeling software generally defaults the base to the largest exposure group. However, the base can be changed by the modeler. The choice of

the base will not change the final predicted values, but it will have an impact on the optics of the factors.

We added an additional parameter while reducing deviance by 149,000. Additionally, the Chi-square statistic is very close to 0%, giving us a high level of confidence that the model with Gender is statistically different than the Intercept Only Model. The formula developed by this model is as follows:

$$E[Y_i] = e^{\beta_0 * x_{0,i} + \beta_1 * x_{1,i}}$$

Where

$\beta_0 * x_{0,i} = Intercept = 1.1617 * 1$ $\beta_1 * x_{1,i} =$

$\begin{cases} 0.1034 * 0 \; when \; Gender = Male \\ 0.1034 * 1 \; when \; Gender = Female \end{cases}$

Therefore, the

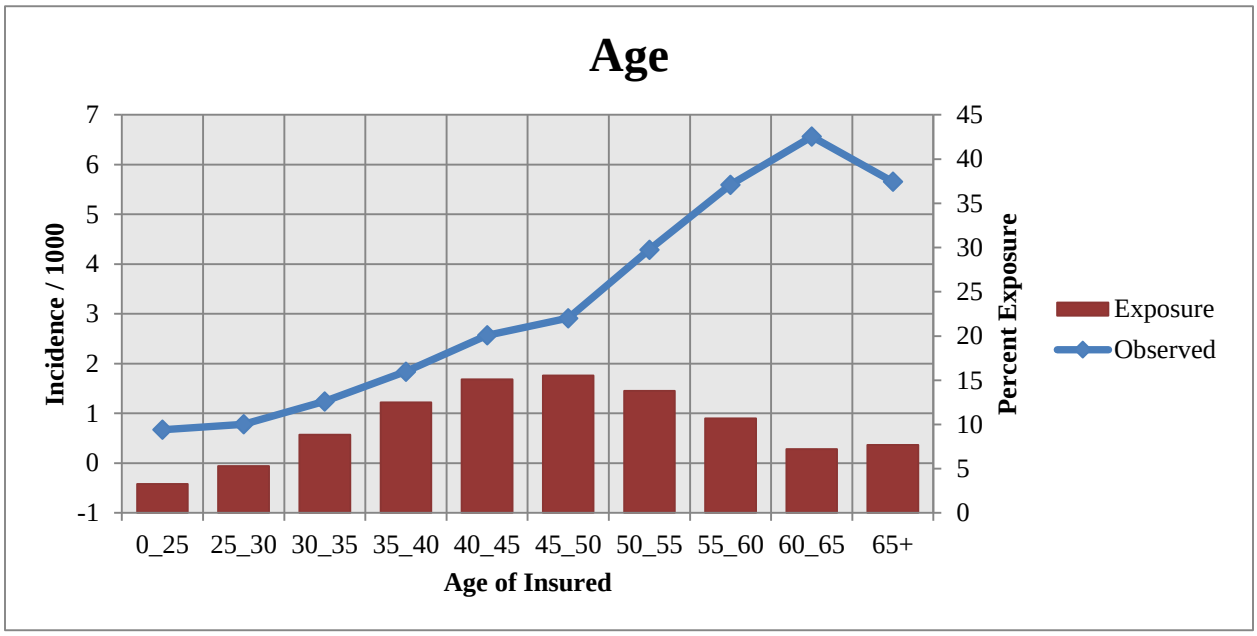
$$E[Male] = e^{1.1617 + 0} = 3.20 \text{ x } 1.00 = 3.20$$

And

$$E[Female] = e^{1.1617 + .1034} = 3.20 \times 1.11 = 3.54$$

The log link function has the added benefit of giving clear relativities between variables. In this case, females are 11% more likely to be disabled than males.

Let us move on to adding another variable – age. During the data stage, we created 11 different age buckets to match the prior rating structure. Segmenting the age variable into two year age buckets would have substantially improved the model fit, but it was not worth the implementation complexity. This is a good example of a practical decision made for easier implementation. The one-way analysis is detailed below:



The table below shows the model before and after adding the age variable.

	Model w/ Gender Only	Model w/ Age	Change
DoF	1,402,895	1,402,886	(9)
Parameters	2	11	9
Deviance	641,878,500	625,147,900	(16,730,600)
AIC	(23,523,740)	(40,254,360)	(16,730,620)
Chi-square			0.00%

The base level is now a male in age bucket 45-50. The vector of parameters and observations is as follows:

Parameters where base is Male age 45-50:

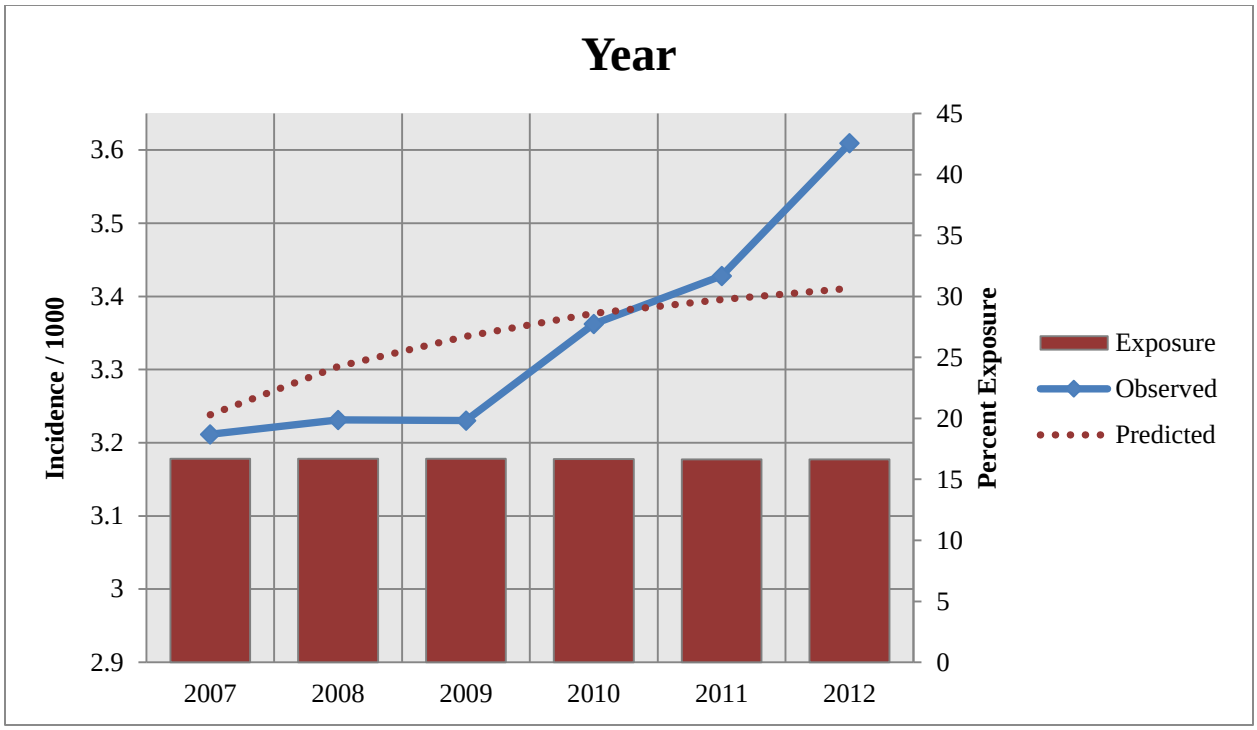
Parameter	Label	Linear Predictor
β0	Intercept	1.0241
β1	Female	0.1025
β2	0_25	-1.4671
β3	25_30	-1.3152
β4	30_35	-0.8535
β5	35_40	-0.4593
β6	40_45	-0.1253
β7	50_55	0.3873
β8	55_60	0.6525
β9	60_65	0.8116
β10	65+	0.6625

The matrix of observations, where the first observation is a male age 45-50 and the second observation is a female age 55-60 would be as follows:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{0,n} & x_{1,n} & x_{2,n} & x_{3,n} & x_{4,n} & x_{5,n} & x_{6,n} & x_{7,n} & x_{8,n} & x_{9,n} & x_{10,n} \end{bmatrix}$$

The 10 columns in the matrix represent the 10 parameters in the above table. The first column is the intercept, which applies to every observation. The second column is the female parameter, which applies to the second observation, but not the first observation.

After adding each variable it is often beneficial to look at the impact on other variables. This will give you a better understanding of your data. For example, if you look at the actual vs. predicted values by year, it is clear that the predicted values increase year-after-year.



Given that only age and gender are in the model and incidence is greater for older individuals and females, the increasing trend implies that the block is aging, increasing the percentage of females, or a mix of both. You can test this theory by removing the gender variable or removing the age variable. Since the predicted pattern is flat without the age variable, it is clear that the book of business is aging.

For each variable, a modeler will first determine if it is statistically significant and then check the impact on other variables. The order of adding variables will vary from modeler to modeler. Although a modeler will generally start with the most important variables (e.g. fit age first when reviewing mortality), individual preference will dictate the order.

Grouping Main Effects

The simple factor model results in 86 parameters. The simple factor model is almost always overfitting, so a modeler must now go through the process of simplifying the model without significantly reducing the model fit.

Significance of Levels

Although a variable may pass the inclusion criteria above, not every level of the variable will be significant. The first step after fitting all the main effects is checking the standard errors for the level of each variable. The standard error percent is calculated by dividing the standard error by the linear predictor. The following criteria are rules of thumb and may be defined differently from modeler to modeler.

Any variable that has a standard error percent greater than 75% is not deemed statistically different than the base. A standard error percent between 50% and 75% is marginally significant. A standard error percent less than 50% is defined as statistically significant. The table below details the standard errors for an SIC based industry grouping:

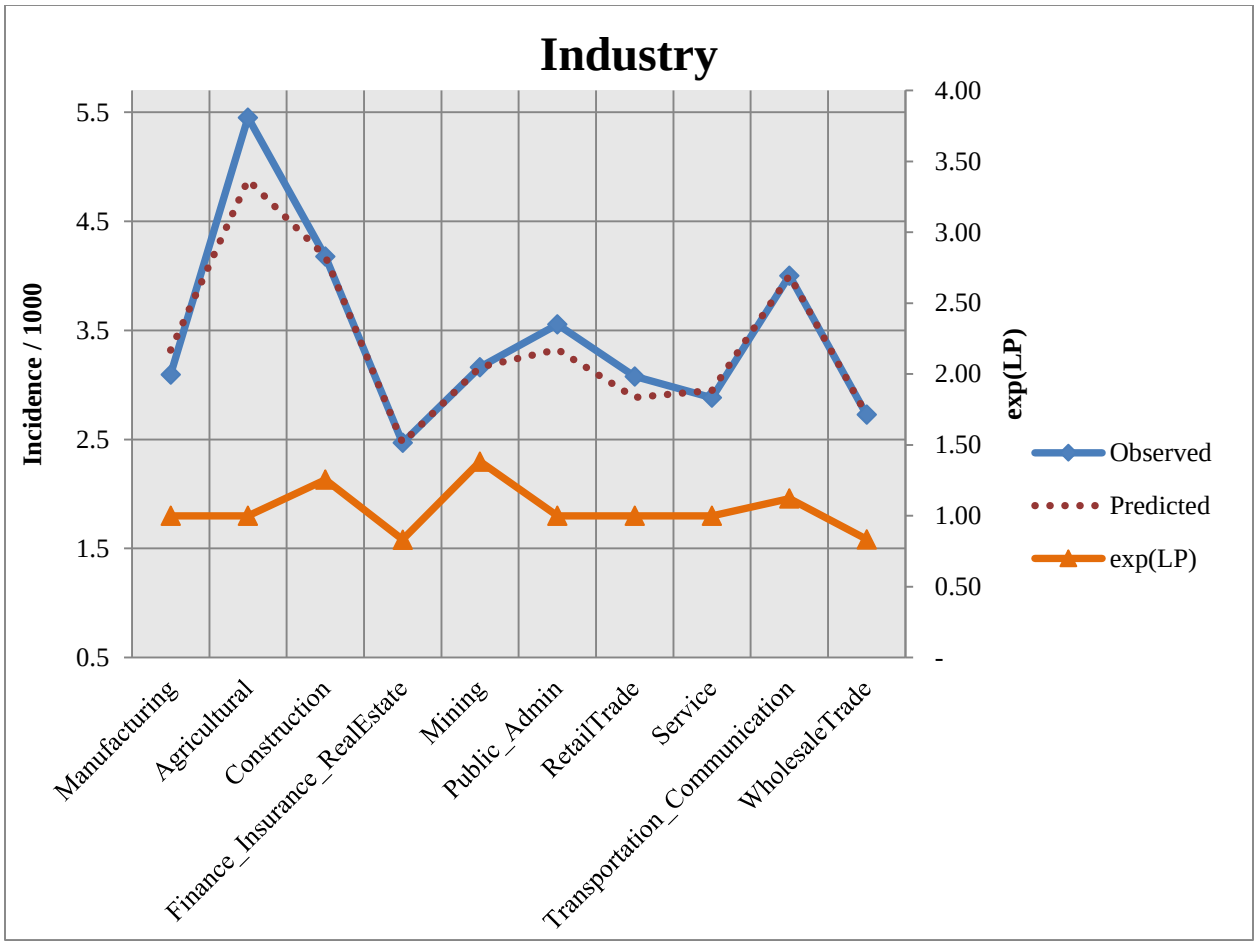
Industry Label	Value	Std. Error	Std. Error %*
Industry_Manufacturing	-0.0453	0.0597	132%
Industry_Agricultural	0.1457	0.0732	50%
Industry_Construction	0.2546	0.0406	16%
Industry_Finance_Insurance_RealEstate	-0.1620	0.0738	46%
Industry_Mining	0.3415	0.0727	21%
Industry_Public_Admin	0.0909	0.0754	83%
Industry_RetailTrade	0.0970	0.0647	67%
Industry_Transportation_Communication	0.1405	0.0542	39%
Industry_WholesaleTrade	-0.1559	0.0709	45%

* Cells above 75% are coded in red and are not significantly different from the base. Cells between 50% and 75% are marginally significant.

In the table above, Manufacturing and Public Administration are not significantly different from the base. Agriculture and Retail Trade are marginally significant. Given that the variables are not significantly different from the base, we chose to group all four of these variables with the base. However, discretion can be used when deciding whether or not to include marginally significant variables. When comparing the model with simple Industry factors to the grouped model in the table below, the Chi-square is above 5% and the deviance increases only slightly. Given that the two models are not statistically different, the best practice is to choose the more simple (i.e. less parameters) grouped industry model.

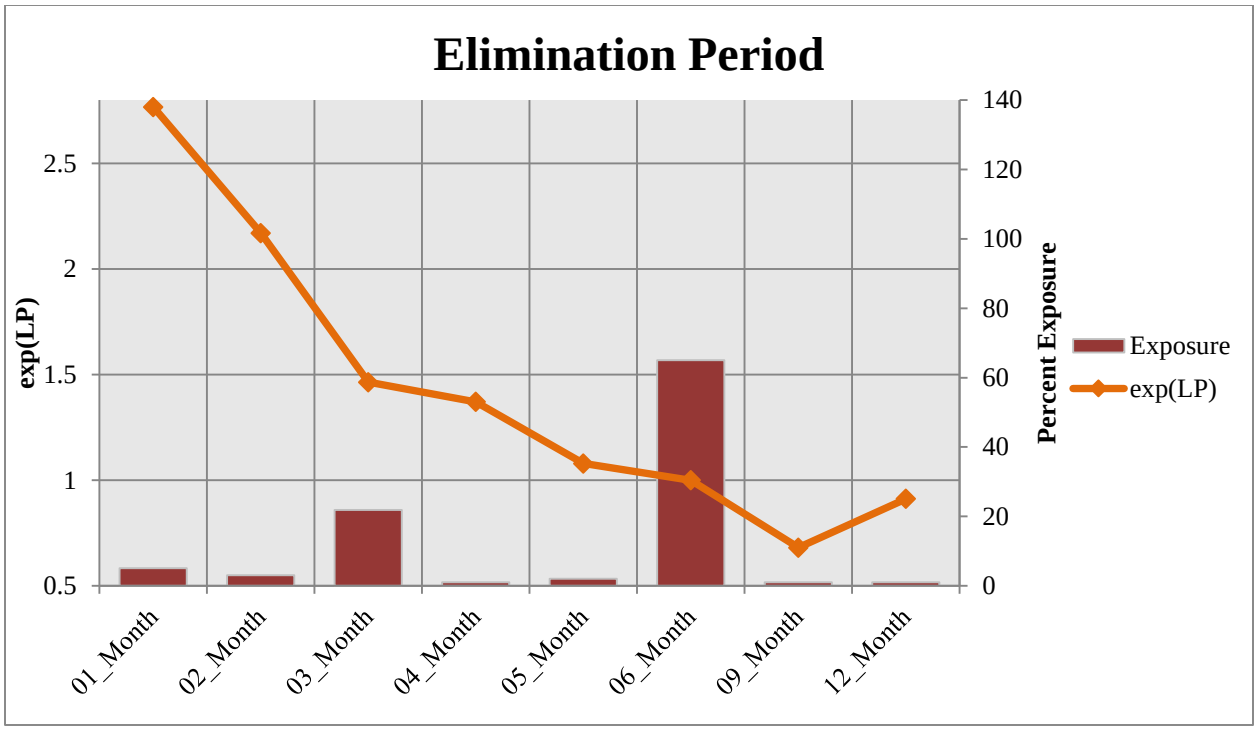
	Simple Industry	Grouped Industry	Change
DoF	1,402,811	1,402,815	4
Parameters	86	82	(4)
Deviance	610,898,000	610,990,300	92,300
AIC	(54,504,140)	(54,411,790)	92,350
Chi-square			10.40%

In the graph below, the exponential of the linear predictors for the previously insignificant variables are now equivalent to the base. The other item of note is that the model still does a good job of predicting the values of the variables that are grouped with the base. For example, Agriculture has an observed incidence that is much higher than the base. Although Agriculture is not included, the model still predicts the higher than average incidence. This implies that other variables are predicting the phenomena such as age, gender, salary, etc.



Counterintuitive Signals

The other instance that requires grouping of simple factors is counterintuitive signals. Below is a graph that shows the exponential of the linear predictor (exp(LP)) for Elimination Period (EP). EP is akin to a deductible in medical or P&C insurance. It is the amount of time an insured must be disabled before they can begin collecting disability benefits. Intuitively, longer EPs (the higher the deductible) lead to lower claim frequency. However, the simplified factor graph below shows a counterintuitive result. We expect a monotonically decreasing slope, but the indications show that a 9 month EP results in lower indicated incidence than the 12 month EP. This could be the result of low credibility or a large exposure that is driving the signal. At this point, the modeler has two choices: 1. dig into the data to determine why this result is occurring (e.g. outlier policy, data error, etc.) or 2. assume that the signal is volatility and account for this volatility by grouping the variable.

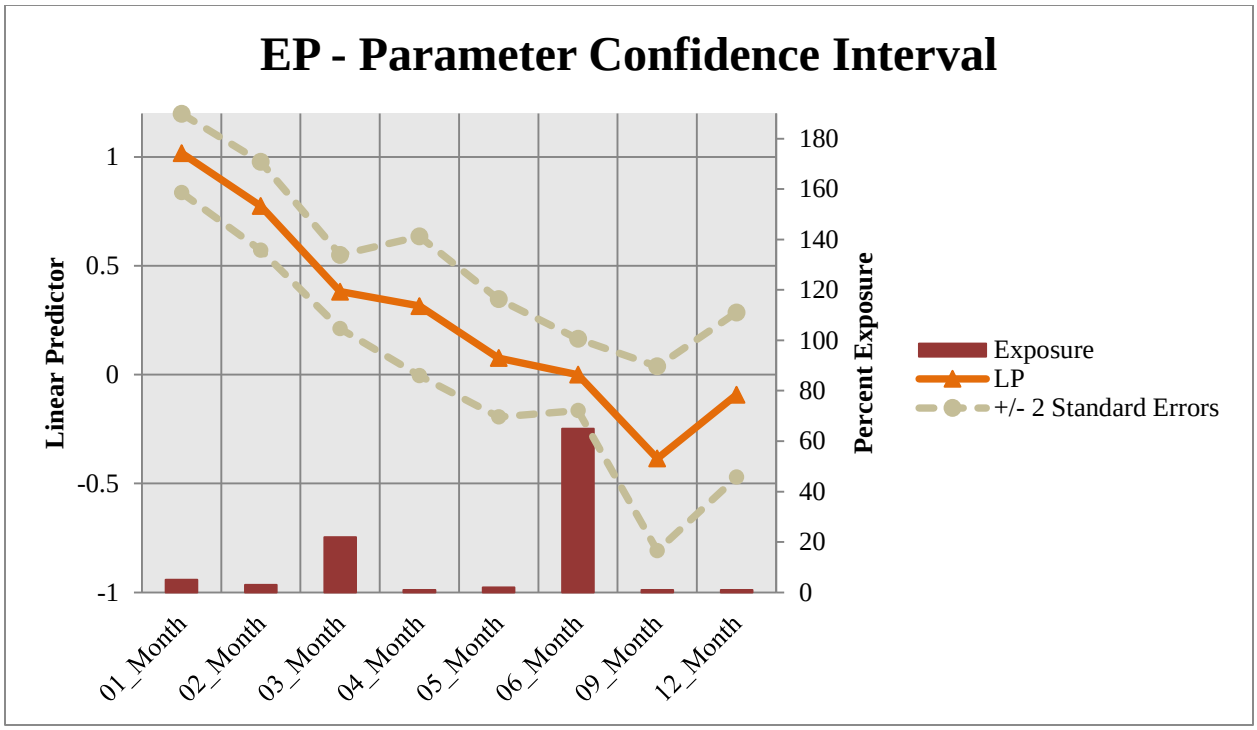


Rather than grouping these variables with the base, it may make sense to group them with the next closest group. For example, group EP 9 and 12 together. Another interesting phenomenon is that the EP 5 indication is not significantly different from the base. The modeler could try to group EP 4 and EP 5 together to increase credibility and come up with an indication that is between EP 3 and EP 6. Performing these modifications resulted in a model that is not significantly different than the simple factor model, but simplifies the model by another 2 parameters.

	Simple EP	Grouped EP	Change
DoF	1,402,815	1,402,817	2
Parameters	82	80	(2)
Deviance	610,990,300	611,027,300	37,000
AIC	(54,411,790)	(54,374,900)	36,890
Chi-square			21.60%

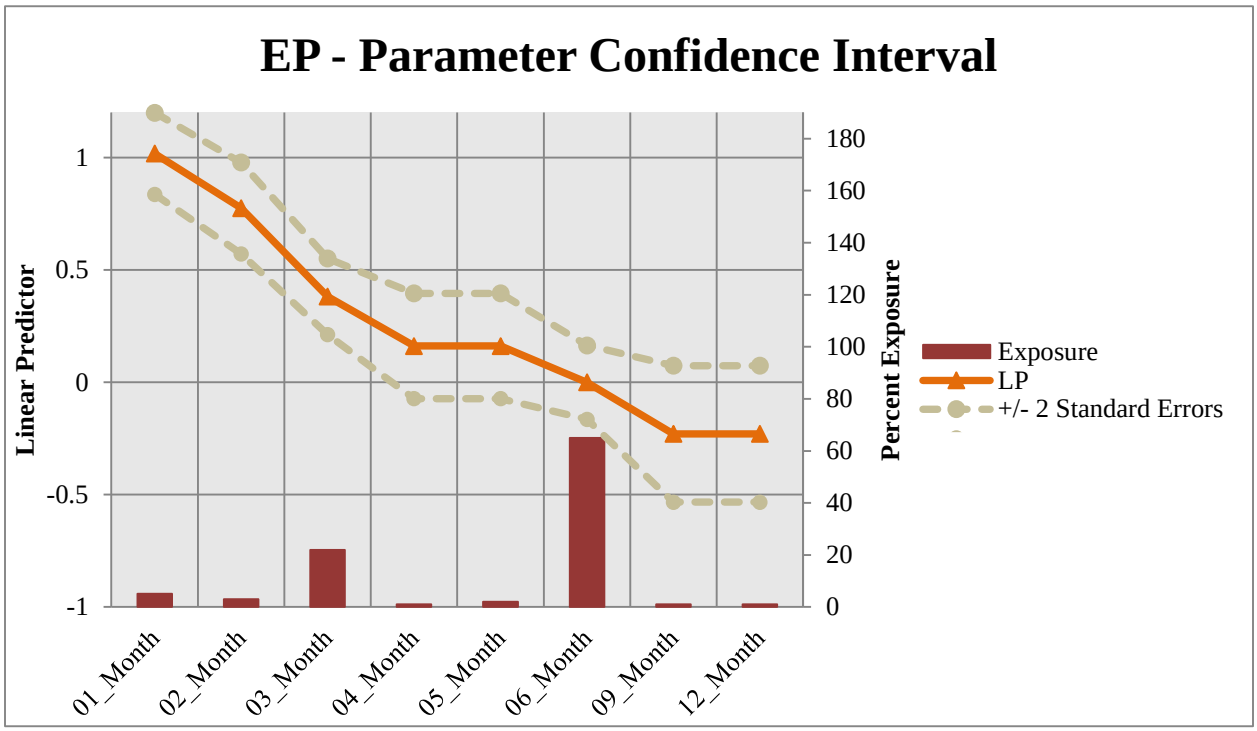
Parameter Confidence Intervals

Another useful technique that blends both the standard error and the counterintuitive discussions above is plotting the confidence intervals of our parameter estimate. In the graph below, we have plotted the linear predictors for EP as well the confidence interval for these parameters. The confidence interval is defined as +/- 2 standard errors and roughly translates to a 95% confidence interval. Said another way, we are 95% confident that the parameter lies within +/- 2 standard errors of the linear predictor.



The 4 and 9 month EPs have very wide confidence intervals. As discussed in the sections above, the only levels that are not significantly different from the base are EP 5 and EP 12. The graph details the same result because the parameters lie within the 6 month confidence interval. When you have a curve that you expect to be monotonically increasing/decreasing, it is also useful to look at the parameter estimates relative to the confidence intervals of the neighbors. For example, the parameter estimate of 9 month is within the confidence interval of the 12 month EP and vice-versa. Therefore, it may make sense to group EP 9 and 12 together rather than group EP 12 to the base. Intuitively, we know that EP 9 and EP 12 have different risks. Since the model cannot distinguish the difference, we will differentiate them in the selection process detailed in subsequent sections.

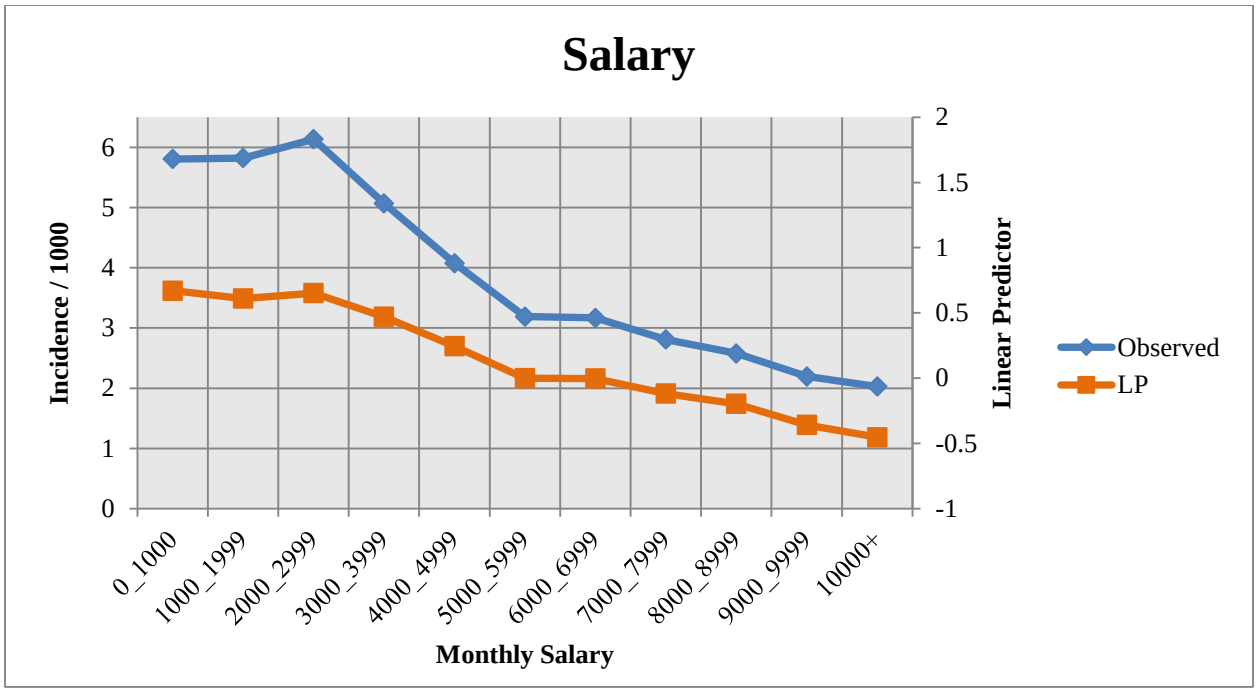
The standard error approach discussed above is useful for determining a variable's significance relative to the base. The confidence interval approach allows a modeler to determine parameter significance relative to neighboring variable levels. After grouping the variables as we discussed above, the graph is as follows:



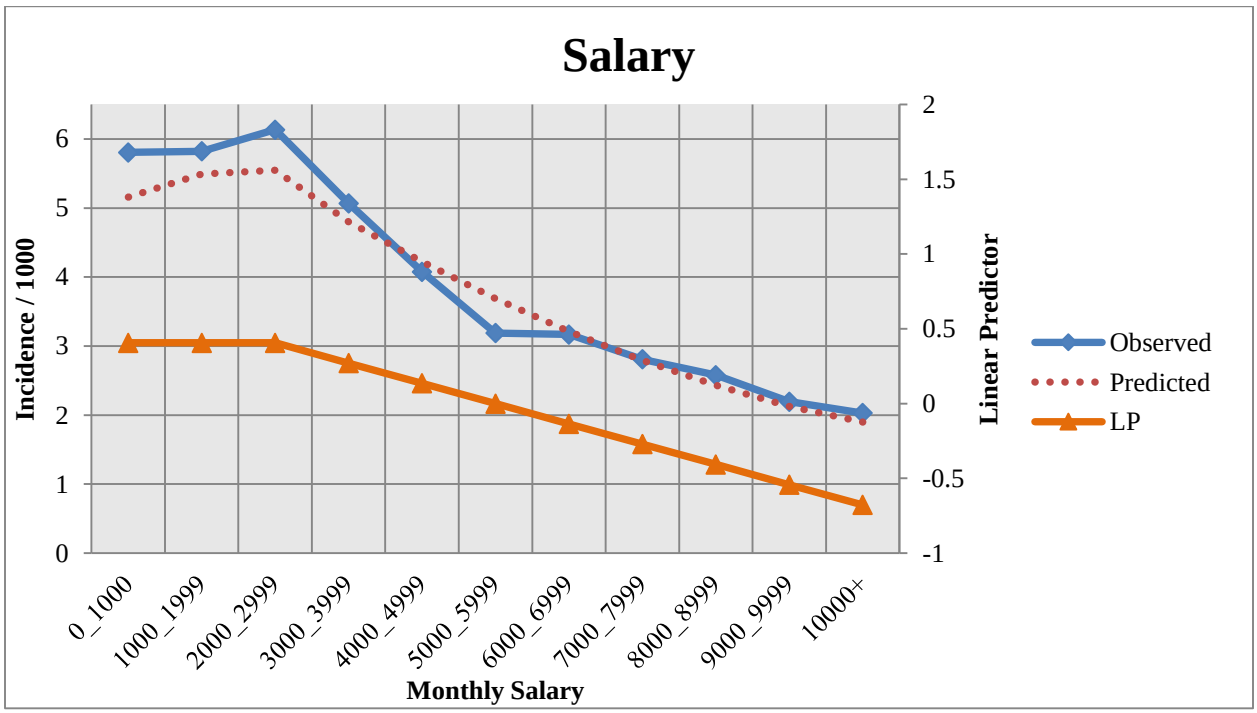
The new grouping ensured that no parameters lie within one another's confidence interval. The linear predictor of EP 4 and EP 5 is at the edge of the confidence interval of both EP 3 and EP 6, indicating that the grouping is marginally significant. However, the curve is monotonically decreasing, which is in line with our expectations. Therefore, this grouping is reasonable and can be used to move forward.

Variates

Variates are another means of smoothing the main effects. They are useful when the variable in question follows a line or curve because they force the structure of the parameters to follow an n-degree polynomial. Consider the example of monthly salary below. From the observed values and the simple factor indications, it is clear that incidence decreases as salary increases. When fitting simple factors, the indications jump around. The salary grouping of 5,000 – 5,999 has effectively the same indication as 6,000 – 6,999.



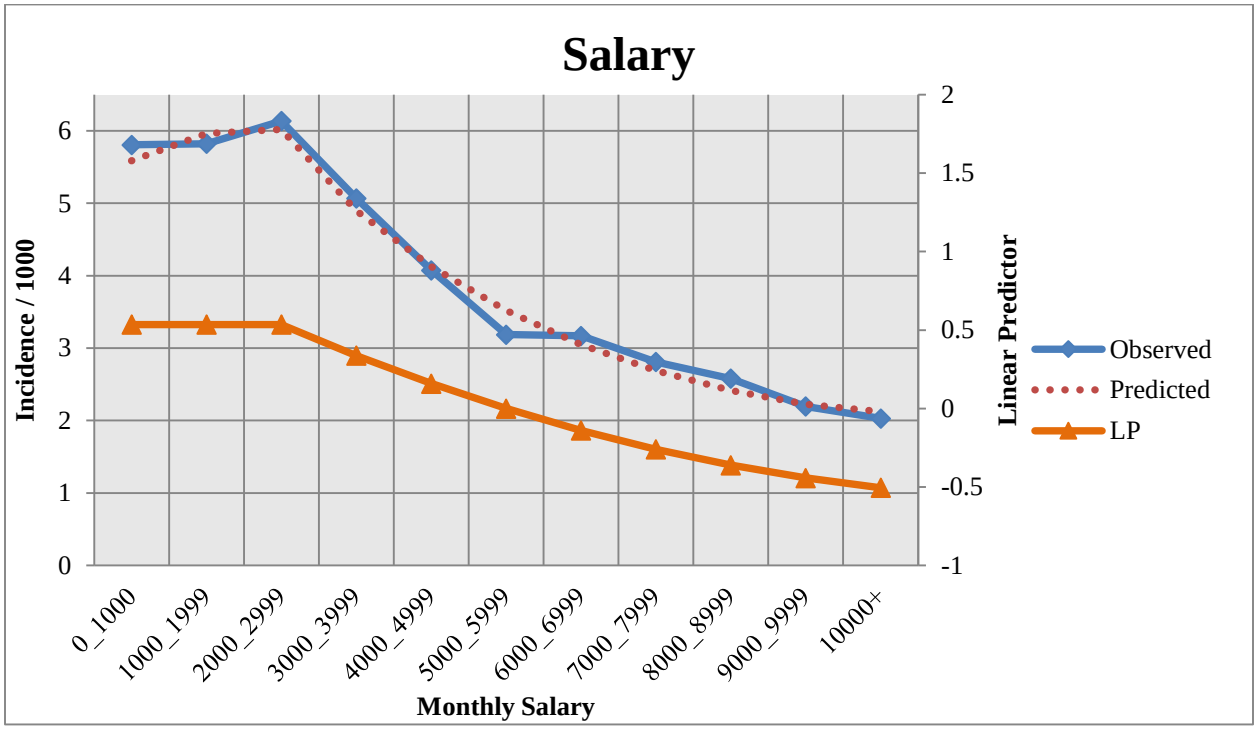
If volatility is the suspected cause, we can smooth the indications by introducing a polynomial. The graph below shows the observed and fitted values after introducing a 1-degree polynomial. The secondary y-axis displays the linear predictors; the linear predictors are perfectly linear. The polynomial could have been fit across all salary bands, but we decided to group salaries of 0 – 3,000 together because the observed trend leveled off at these salary buckets.



The introduction of these variables yielded the following statistical results:

	Simple Factors	1° Polynomial	Change
DoF	1,402,817	1,402,826	9
Parameters	80	71	(9)
Deviance	611,027,300	611,336,600	309,300
AIC	(54,374,900)	(54,065,600)	309,300
Chi-square			0.20%

Adding a 1-degree polynomial reduced the number of parameters by 9, making the model simpler. However, a deviance increase of 309,300 implies that the model fit is worse. This is evident when you compare the actual and predicted values. We are under-predicting the end points and over-predicting the middle salaries. A Chi-square of 0% indicates that the two models are statistically different. Therefore, we have simplified the model at the expense of predictive power. When the 1-degree polynomial is not sufficient, we can try a 2-degree polynomial.



The graph above shows that the linear predictors no longer follow a linear trend. They are decreasing at a decreasing rate. This allows the model to better fit the lower and upper salary ranges. When we look at the comparison of the simple factor model with the 2-degree polynomial model, we now see a Chi-square statistic above 5%. Since 5% is our threshold for statistical significance, we can now say that the two models are not significantly different from one another. Therefore, we were able to reduce the number of parameters by eight without significantly reducing the model fit.

	Simple Factors	2° Polynomial	Change
DoF	1,402,817	1,402,825	8
Parameters	80	72	(8)
Deviance	611,027,300	611,169,800	142,500
AIC	(54,374,900)	(54,232,360)	142,540
Chi-square			15.80%

Mid-Model Grouping

You may have noticed that there are a large number of variables in the simple factor model (86). Forty nine of those parameters come from the state level. Most of the state parameters are not statistically significant. Therefore, a modeler may decide to create a new group in the middle of the modeling process. In this instance, we mapped the state variable to the nine regions that are defined by the Bureau of Labor Statistics. The statistics below show that the region grouping reduces the number of parameters by 41. The Chi-square statistic of 17.1% signifies that the model with region is not significantly different than the model with state.

	State Model	Region Model	Change
DoF	1,402,825	1,402,866	41
Parameters	72	31	(41)
Deviance	611,169,900	611,764,900	595,000
AIC	(54,232,200)	(53,637,350)	594,850
Chi-square			17.10%

The large increase in deviance occurs because we have removed a large number of parameters from the model. Since we no longer have a parameter for each individual state, we expect a large deviance increase (i.e. the predicted values are not as close to the actuals). Using the state level variable is a good example of overfitting. Although the predicted values would be closer to the observed (i.e. lower deviance), the model with 41 additional parameters would not fit future data as well.

As we have seen in the examples above, grouped factors and variates are added to create the main effects model. Once finished, the number of parameters was reduced from 86 in the simple factor model to 23 in the main effect model.

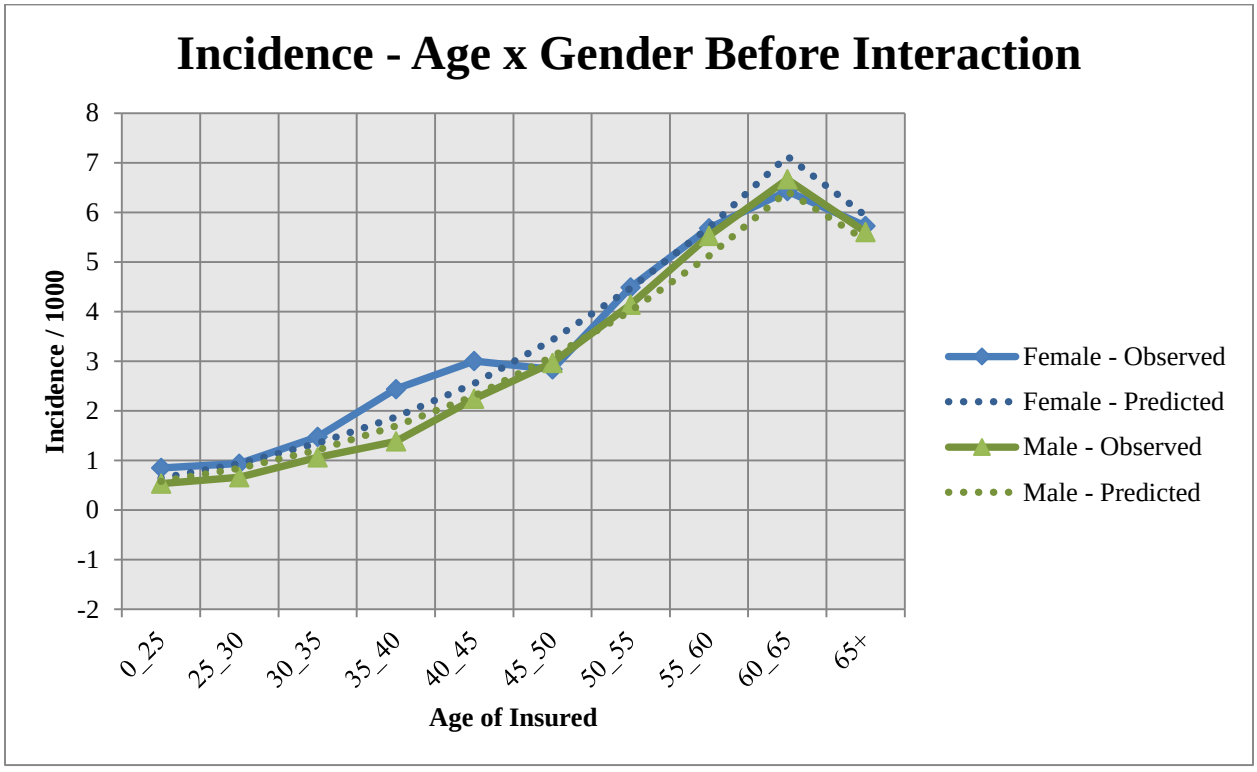
Control Variables

Control variables are modeled in the same manner as all other variables. The greatest distinction between the two is that a control variable is included to control for a phenomena that we do not expect to continue in the future. Year is the most common type of control variable. A good example is disability rates during a recession. Since disability income insurance claims occur more often when the prospect of losing one's job is greater, carriers often experience increased disability incidence rates in times of economic stress. If the most recent year in the data is the base, and all prior years had worse than expected disability incidence, the introduction of the

year variable would show linear predictors above zero for years of economic distress and a decreased intercept value. When implementing the model, this new intercept would better reflect the expected future state of incidence.

Interactions

Interactions allow the impact of one variable on another to be explained by the model. As seen in the earlier example, females are on average more likely to be disabled than males. Is this likelihood consistent across all other variables? It is well known in the disability insurance industry that the difference in incidence between females and males is much greater at younger ages than older ages. This stems from women of child bearing ages having increased likelihood of disability. The graph below is the result of the main effect model and illustrates the actual versus predicted incidence for males and females across all age buckets. The model under-predicts the incidence of younger females and over-predicts the incidence of younger males.



Adding an interaction is similar to adding parameters as we discussed previously. The only difference is that these new variables are identified by the intersection of two other rating variables. Interacting the simple factors for age and gender results in the following:

	Main Effect Model	Model w/ Interaction	Change
DoF	1,402,874	1,402,865	(9)
Parameters	23	32	9
Deviance	610,198,400	609,661,900	(536,500)
AIC	(55,203,830)	(55,740,330)	(536,500)
Chi-square			0.00%

The criteria for adding interactions are the same as adding main effects. The model statistics above indicate that the interaction is statistically significant. The additional 9 parameters and the standard errors are as follows:

Parameter	Label	Value	Std. Error	Std. Error %
β_{24}	Female & 0_25	0.7431	0.2928	39%
β_{25}	Female & 25_30	0.3914	0.2206	56%
β_{26}	Female & 30_35	0.4257	0.1440	34%
β_{27}	Female & 35_40	0.5437	0.1009	19%
β_{28}	Female & 40_45	0.3939	0.0868	22%
β_{29}	Female & 50_55	0.1571	0.0811	52%
β_{30}	Female & 55_60	0.1235	0.0861	70%
β_{31}	Female & 60_65	0.0022	0.1028	4672%
β_{32}	Female & 65+	0.1070	0.1062	99%

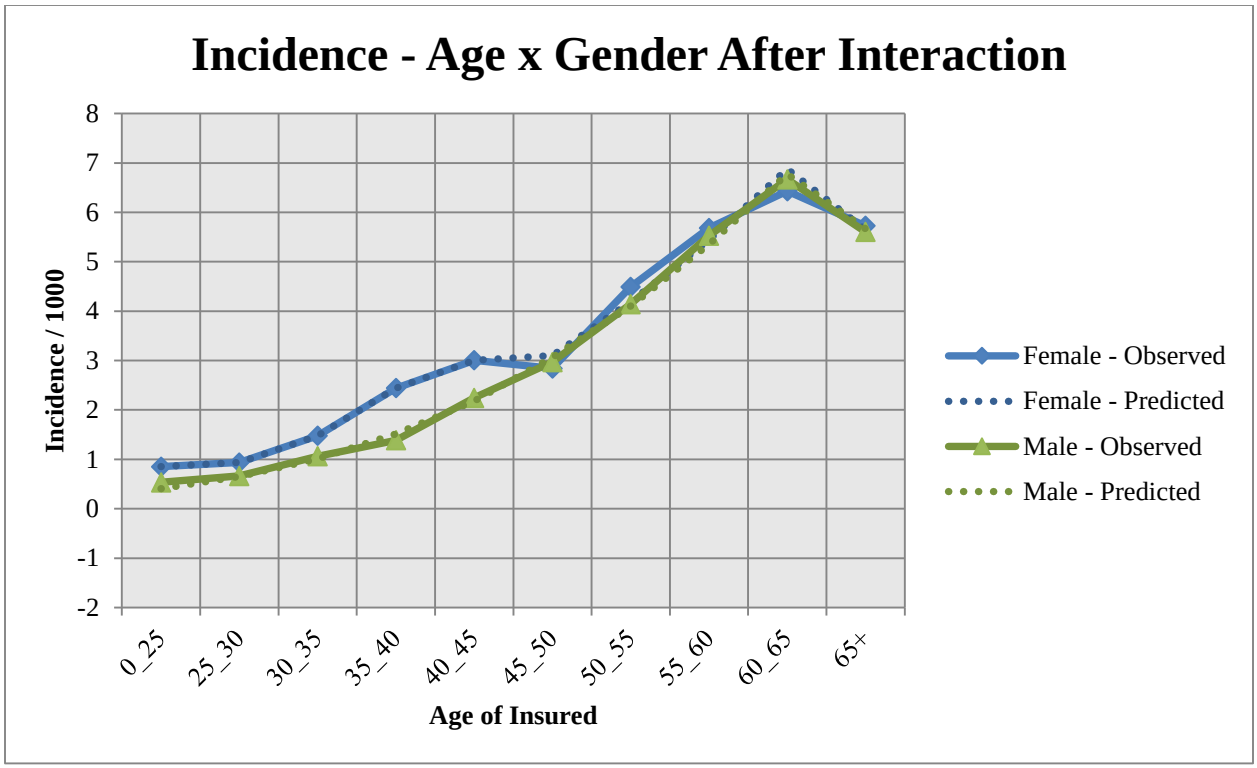
The table to the left has no parameter for the 45-50 year old female. This bucket is defined by the main effect female factor. The standard errors show that the interaction above age 50 is not statistically significant. Therefore, you would simplify the model so that every female older than 45 would be defined by the female main effect.

It is recommended that a modeler be more conservative when adding interactions. Interactions should only be included in the model when they are both statistically significant and make intuitive business sense. Interactions are a more granular view of the data compared to the main effects, so the credibility will naturally be lower. Given this lack of credibility, interactions that are statistically significant may, in fact, be overfitting. For this reason, it is wise to remove marginally significant interaction variables from the model unless well justified for business reasons.

We decided to remove the parameters above 50, but left the 25-30 parameter for the time being. Knowledge about younger females exhibiting higher than average incidence led to this conclusion. The exhibit below shows that the model with the full interaction is not statistically different from the simplified interaction. This implies that the difference in incidence between males and females above age 45 does not vary by age.

	Full Interaction	Interaction Below Age 45	Change
DoF	1,402,865	1,402,869	4
Parameters	32	28	(4)
Deviance	609,661,900	609,736,200	74,300
AIC	(55,740,330)	(55,666,040)	74,290
Chi-square			18.70%

When reviewing the actual versus predicted, the model with the interaction below age 45 does a better job predicting incidence by age and gender.



After adding an interaction, it is beneficial to review the change in main effects. The exponential of the female linear predictor was 1.11 before adding the interaction. After adding the interaction, this value reduced to 1.02. This implies that the overall differential in incidence between males and females is driven purely by those females below age 45. When reviewing the standard errors, the female parameter has a standard error percent well above 100. In this instance, we can make the decision to either keep or remove the main effect. We chose to keep the main effect, although there is no statistical rationale. If the model structure were to be applied to a different dataset in the future, we would not want to ignore a potential difference in incidence at the older ages.

After adding interactions, the standard errors should be reviewed again and any variables that are no longer significant may be removed. We will proceed to validation with a model that has 28 parameters.

Model Validation

Once you are satisfied with the model built on the train dataset, it is time to check how the model performs on the test dataset. There are three methods of validation that can be performed:

Parameterize Model on the Test Dataset

After you adapt the model on the test dataset, you should review the signals to check for signals that do not align with the train model. Additionally, the test dataset standard errors should be

reviewed. In the exhibit below, we detail the standard errors for the industry variable and the age/gender interaction. The table below shows that the Wholesale Trade and Transportation/Communication industries and the Female/Age 0-25 variables were both significant when parameterized on the train dataset. However, on the test dataset, these variables are insignificant.

	Train			Test		
Construction	0.2261	0.03599	16%	0.3238	0.04339	13%
Finance_Insurance_RealEstate	-0.1869	0.07258	39%	-0.2083	0.08995	43%
Mining	0.3143	0.07111	23%	0.4964	0.07999	16%
Transportation_Communication	0.1132	0.05137	45%	0.0409	0.06683	163%
WholesaleTrade	-0.1814	0.06904	38%	-0.0267	0.07985	299%
Female & A_0_25	0.7181	0.28235	39%	0.5924	0.38183	64%
Female & B_25_30	0.345	0.20726	60%	0.7181	0.23088	32%
Female & C_30_35	0.361	0.13274	37%	0.4798	0.15729	33%
Female & D_35_40	0.4652	0.08812	19%	0.5008	0.1059	21%
Female & E_40_45	0.3081	0.07112	23%	0.3233	0.08456	26%

This example details one of the major strengths of predictive modeling. The train/test validation allows identification of these false signals. Therefore, these parameters can be removed (i.e. grouped with the base level).

After reviewing the other variables, the model parameters were reduced to 24 from 28.

Compare Train and Test Factors

Another means of determining false signals is comparing the parameters from the train and test datasets. This allows determination of whether the parameters are stable between the two independent datasets. Below, we compare the train and test factors for EP and the Age/Gender Interaction.

	exp(LP)		
	Train	Test	Abs(Diff)
01_Month	2.6464	2.6894	2%
02_Month	2.0913	1.7545	19%
03_Month	1.3792	1.3967	1%
04_Month	1.3792	1.3967	1%
05_Month	1.0000	1.0000	0%
06_Month	1.0000	1.0000	0%
09_Month	0.7245	0.7678	6%
12_Month	0.7245	0.7678	6%
Female & B_25_30	1.2652	1.8815	33%
Female & C_30_35	1.3384	1.5315	13%
Female & D_35_40	1.5329	1.6027	4%
Female & E_40_45	1.3398	1.3656	2%

When performing the factor comparison, it is important to be on the lookout for two items: factors switching signs and factor instability.

Switching signs occurs when the exponential of the linear predictor is above/below 0 in one dataset and below/above 0 in another dataset. Since the variables direction versus the base level cannot be determined, these variables should be grouped with the base.

Factor instability occurs when the direction versus the base is the same between train and test, but the difference between the exponential of linear predictors is large. The exhibit shows two variable levels that show instability. The 2 month EP has a 19% difference between train and test. However, the indications in both datasets lie between the 1 and 3 month EP indications. Therefore, judgment can be used to determine how to treat these variables.

When reviewing models that do not use a log link function, you can compare linear predictors to determine if the variable is switching signs. To compare factor stability, you may want to apply the intercept and the parameter being reviewed to the link function. For example, in a logit link model with a binomial error term, you would calculate the following:

$$\frac{e^{Intercept+Linear\ Predictor}}{1+\ e^{Intercept+Linear\ Predictor}}$$

This will allow the modeler to compare the two datasets and the effect of the linear predictor on the final formula.

The level of the interaction for Female Ages 25-30 and Female Ages 30-35 show another instance of factor instability. Although the train and test indications are both greater than 1, their relativity to the next grouping (ages 30-35) changes direction. In the train dataset, the 25-30 indication is less than the 30-35 while in the test dataset the 25-30 indication is greater than the

30-35 indication. You must decide whether the instability justifies grouping the 25-30 level with the base or if it is appropriate to group 25-30 and 30-35.

In this instance, we chose to group 25-30, 30-35, and 35-40. Business intuition would imply that women in these age groups experience higher than average long term disability incidence. Since females age 25-30 fall in this category, it was more appropriate to group them with ages 30-35 than with the base.

This process can be repeated for multiple versions of train and test to gain comfort in your model. As mentioned earlier, you can compare factors from datasets in two different time periods (eg. 2007-2009 vs 2010-2012) or datasets from different sampling methods.

After making these adjustments, the model had 22 parameters.

Offset Train Model and Score on the Test Dataset

If the modeling software allows the train model to be scored on the test dataset with ease, it is a worthwhile exercise. The results of this analysis will be similar to factor validation above for variables that are included in the model. However, this approach is more useful when reviewing the models performance on variables that are excluded from the model. Offsetting refers to the approach of fixing model factors. Scoring is the approach of applying these factors to a new dataset. In other words, you are just applying the algorithm to the observations in the test dataset. It is useful to look at the actual versus predicted (A/P) values to determine if the model is sufficient. An experienced modeler expects the A/P to be very close for variable levels with sufficient credibility. We expect greater A/P deviation for levels with low credibility. A good rule of thumb for evaluating model sufficiency is if the actual versus predicted values are within 5%. A model can still be sufficient if the A/Ps are greater than 5%. The modeler and business partners must determine this threshold based on the context of the problem being solved and business intuition.

Backwards Regression

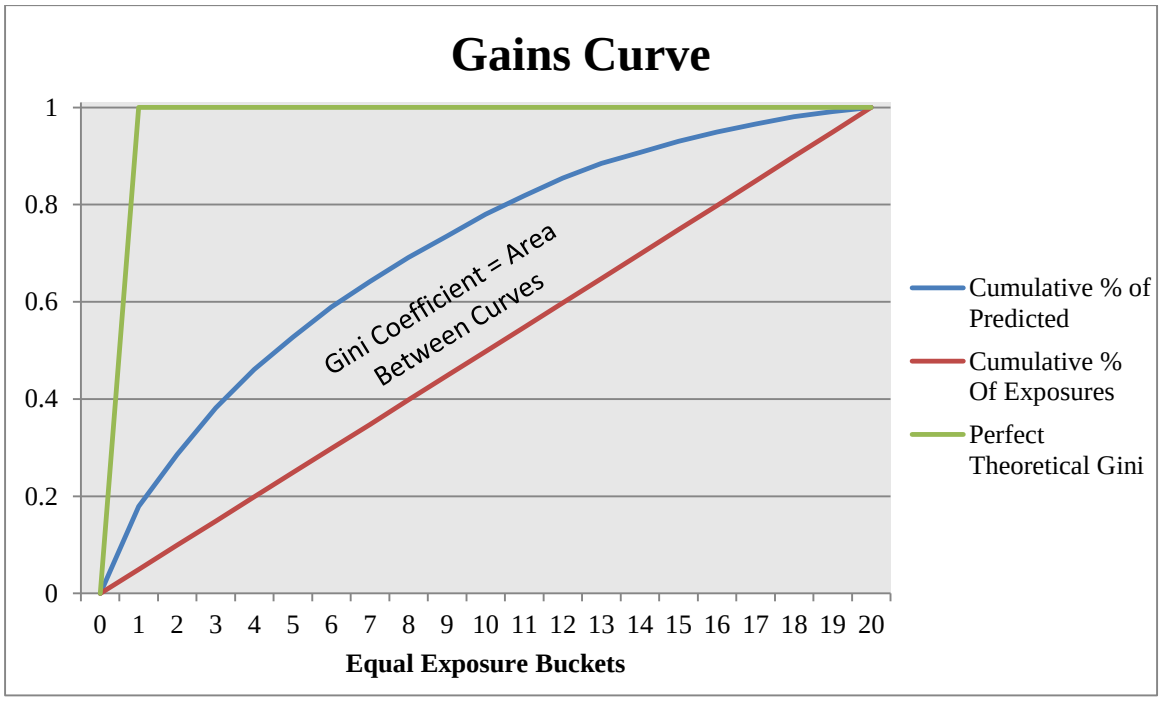
Backwards regression is a technique that reverses the model building process. We added variables to the model to check their significance, while backwards regression removes variables to test their significance. After each variable is removed, the same statistics will be analyzed (Chi-square, AIC, etc.). If a variable fails the criteria (e.g. Chi-square > 5%), the variable will be removed. If the variable passes the criteria, it will be added back to the model and the next variable will be tested. This continues until all variables in the model have been tested. It is prudent to perform this test on both the train and test datasets.

The final validated model has 22 parameters – a reduction from 86!

Model Evaluation

Gini

The Gini score is a metric that signifies how well the model segments the response. The metric is calculated by ordering the predicted values from highest to lowest. The data is then bucketed into equal exposures. Starting with the bucket with the highest predicted value, the cumulative sum of predicted values is calculated and plotted with the cumulative percentage of data. The Gini coefficient is the area between the two curves. The higher the Gini value, the better the model segments the variables.



The Gini value is difficult to compare across projects. It is more appropriate to compare the Gini scores across two models built on the same data. It is not appropriate to compare the Gini scores across two separate datasets.

The green line represents a perfect Gini score. This is only possible if less than 5% of your exposures represent the entire value of the response (e.g. all claim dollars or all claim counts). If the response has positive values across a large portion of your exposures, the maximum Gini score is reduced.

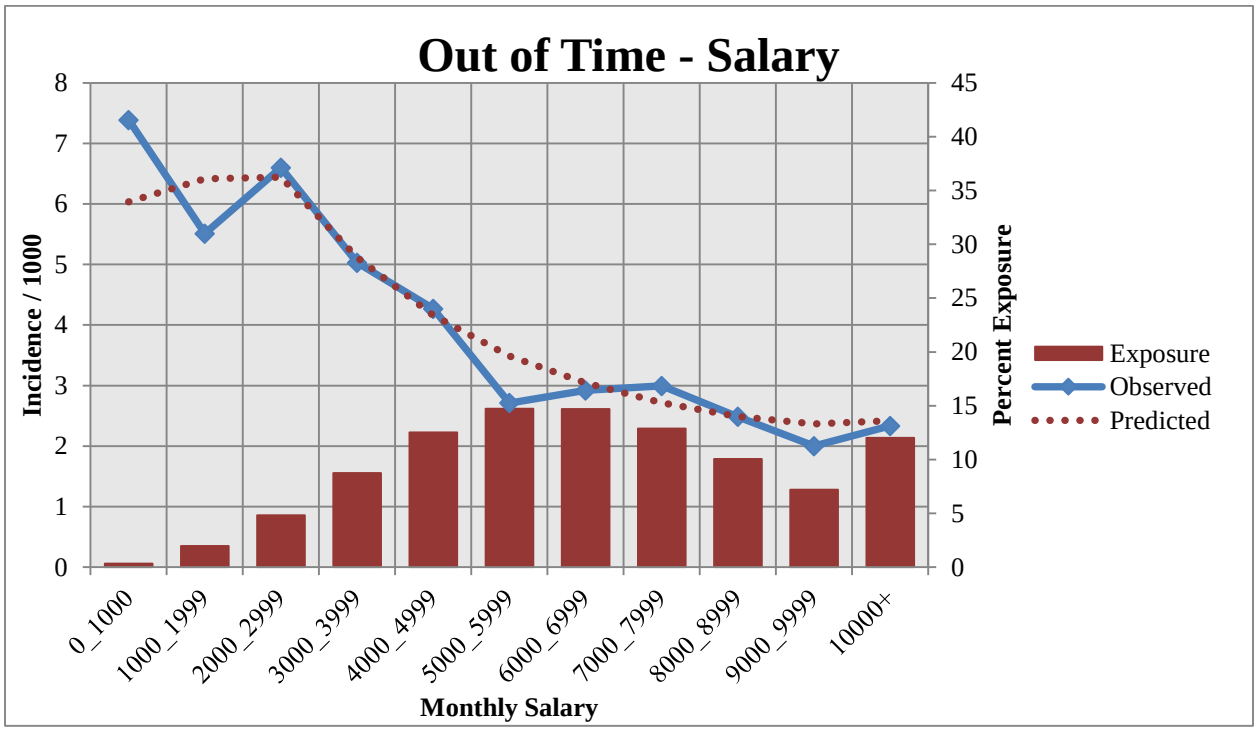
Out-of-Time

As mentioned earlier, the 2013 data was reserved for out-of-time sampling. This test is very similar to the offset train and score on test approach above. In this test, you take the model developed above and score it on the withheld data. The first check is to determine if the model predicts the response on average. The comparison of incidence is as follows:

	Actual	Predicted	A/P-1
2013	3.32	3.45	-4%

Actual incidence is 4% less than predicted incidence. As long as the model does a good job of segmenting risks (predicting the slope) the slight difference in the overall level is not concerning. The only consideration is what intercept should be used when implementing the model. Do the modeler and business partners believe that the 2013 incidence was a better than average year? If yes, the intercept from the original model may be appropriate. If the belief is that the 2013 incidence is more indicative of the future, a 4% decrease to the intercept may be appropriate.

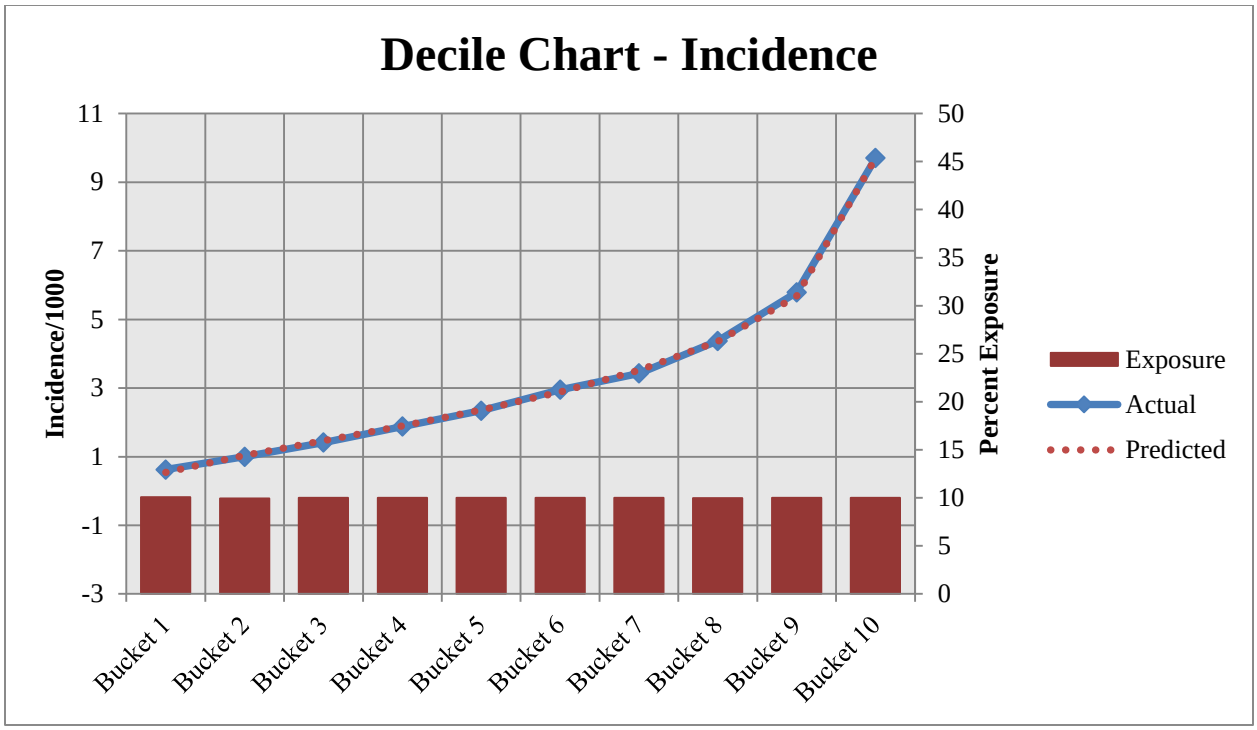
As mentioned above, it is important to review the actual and predicted values of each variable. The graph below reviews the observed and predicted values for salary in the out of time dataset.



The graph shows the actual line jumping around the predicted values. Given that this is only 1 year of data, we do not expect the lines to match exactly. As long as the actual values appear to indicate volatility (actuals jumping around predicted values) rather than an emerging trend (actuals moving further and further from predicted at the endpoints), we feel comfortable with the model.

Decile Charts

A decile chart is produced in a similar manner to the Gini coefficient. The model is applied to each observation and sorted from lowest predicted value to highest predicted values. Ten buckets are then created with equal exposures. You then compare the actual and expected values in each bucket. The graph below shows that the actual and predicted values align in each bucket.

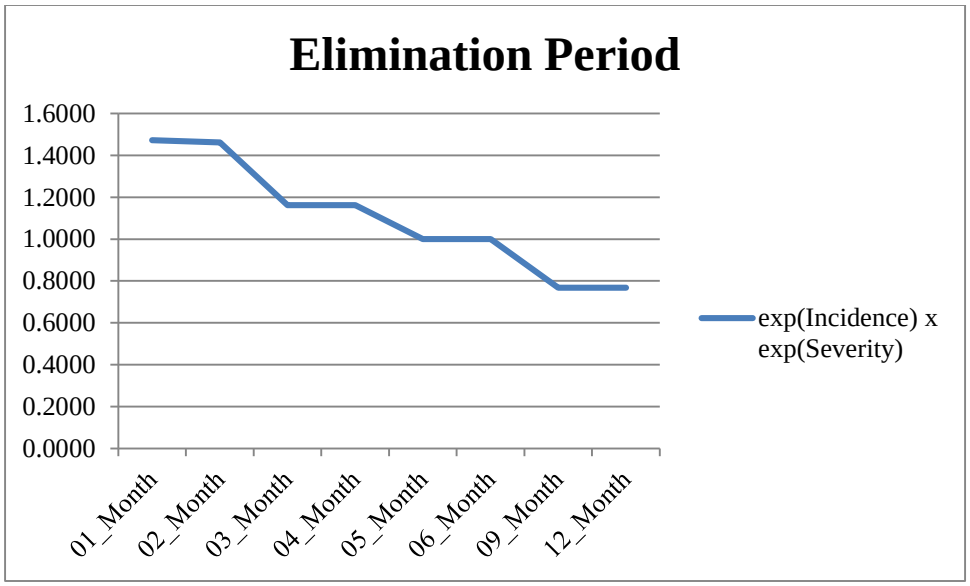


You may run into instances where the A/P is not close and you can investigate the particular bucket to determine if there are specific cases that are not aligning. It is also useful to perform the decile analysis on other versions of the data. For example, you can calculate the graph on train, test, and the out-of-time sample.

Combining Models

Once the aforementioned process has been completed for both the incidence and severity models, it is now time to combine them to develop the total estimated claim cost. At this point, the choice to use the log link for severity in lieu of the canonical link function becomes clear. The log link is multiplicative, therefore combining the factors for each variable is as simple as multiplying the incidence and severity factors together. The table and graphs below show the final indicated factors for elimination period and the associated graph:

	Incidence	Severity	***Final Indication***
01_Month	2.6894	0.5475	***1.4726***
02_Month	1.7546	0.8332	***1.4618***
03_Month	1.3968	0.8319	***1.1620***
04_Month	1.3968	0.8319	***1.1620***
05_Month	1.0000	1.0000	***1.0000***
06_Month	1.0000	1.0000	***1.0000***
09_Month	0.7679	1.0000	***0.7679***
12_Month	0.7679	1.0000	***0.7679***



The combined model factors will then be used in the selection process.

Selection and Dislocation

Once you are finished evaluating the model and you are comfortable with the results, it is time to begin the post-modeling phase. This involves selection and dislocation.

Selection

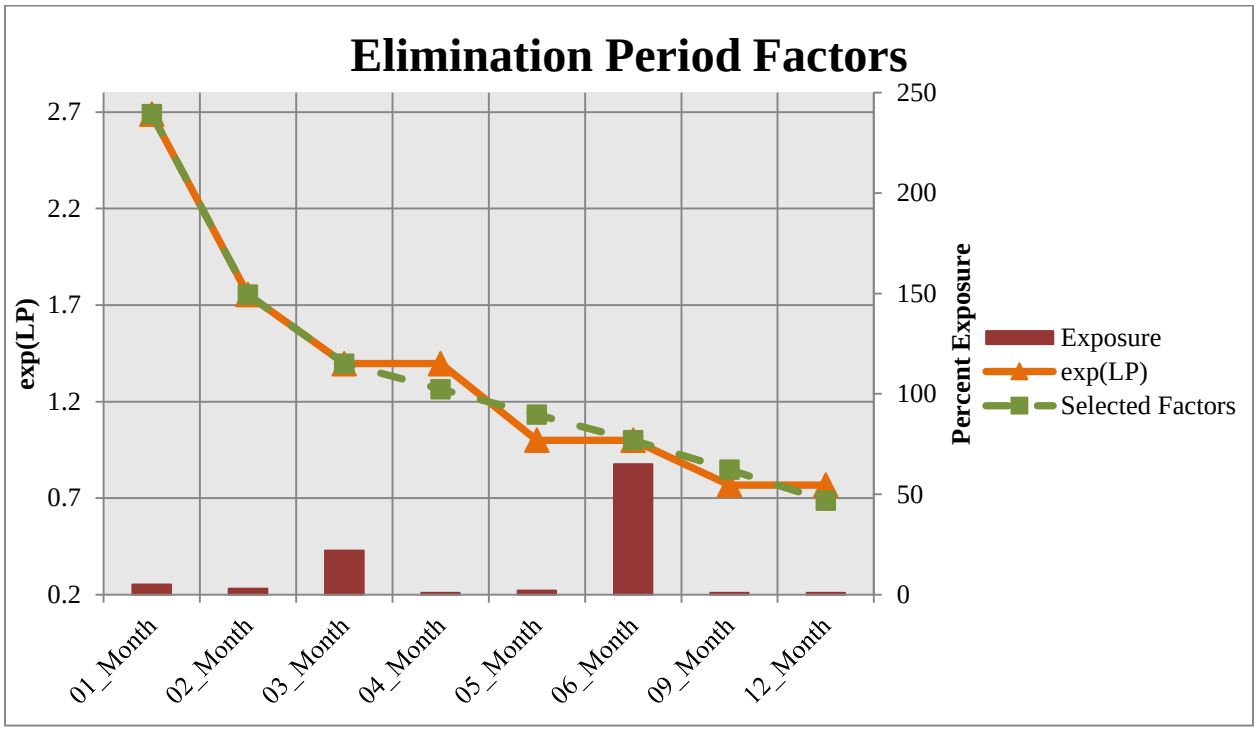
Selection is the process of either developing factors that could not be derived in the modeling process or adjusting the modeled indications.

A successful selection process should involve the expertise of both the modeling team and the business partners. The modelers will be able to provide insights that could only be derived through model building, while the business partners have insights that could only be derived through years of experience.

As we discussed earlier in the model participants section, this may be a good time to bring sales and underwriting into the discussion. The selection process may benefit from their front line knowledge of the product and the potential impact of some of the modeling changes.

In the modeling phase above, we discussed the need to group the Elimination Period variable because not all levels were statistically significant. Although we grouped EP 4 with EP 3 and EP 5 with EP 6 during the modeling phase, we know that this grouping does not make sense as we look to implement the model. In the graph below, we show the modeled factors and the selected factors. EP 4 and EP5 were linearly interpolated values between EP 3 and EP 6. We chose linear interpolation because EP 4 and EP 5 have such low exposures that they would not materially impact the initial indications of EP 3 and EP 6. For EP 9 and EP 12, we chose to

group them together in the modeling phase. Therefore, the grouped indication is somewhere between the true impact of the two variables. Therefore, the selected factor for EP 9 is above the grouped indication and the selected factor for EP 12 is below the grouped indication.



The selection phase is much more of an art than a science. When looking at the results from EP 1 to EP 3, it seems that the graph follows a convex shape. We chose to linearly interpolate, but we could have chosen to follow the convex shape of the graph for the EP 4 and EP 5 selections. The decision is up to the modeling team and their business partners.

Dislocation

After the selection phase is completed, dislocation is the comparison of the newly developed algorithm versus the old approach. During the LTD pricing project, it would be the comparison of old versus new premiums.

To this point, we have only looked at the pieces to construct estimated claim cost. However, to truly see the impact of the model, it is necessary to gross up the claim costs for expenses and profit. After applying expenses and profit, it is possible to look at the impact of the model on the market rates. In this context, dislocation refers to the change in premium across notable variables. For example, sales will be interested in the dislocation of premium across sales region or industry classification.

Implementation

The first phase of implementation is building the scoring engine. The scoring engine is the infrastructure that applies the final algorithm to future data. This is not generally performed by the modeling team or business partners. However, the modeling team will be heavily involved in giving direction to IT. The most important aspect of IT implementation from the business partner's point of view is testing. It will take a considerable amount of time to ensure that the months of work that were invested in building your model are not foiled by implementation mishaps.

After the scoring engine is built, the team will need to determine how the score (predicted values) will be interpreted. This will range from building a structured decision engine to training individuals to interpret the scoring engine. The raw score or the interpretation of the score will then be delivered to the end users through some sort of interface. The scope and structure of the implementation phase will vary from project to project.

Documentation

Often an overlooked aspect of any project, documentation is imperative for a few reasons:

1. Future modelers can use your documentation as a guide to understand the learnings that will not be clear from the pricing algorithm. This will save considerable resources when trying to refresh the model in the future. Refreshing the model is the process of applying new data to the model to see if the indications have changed.
2. Documentation will increase the credibility of the modeling team. Having clear documentation on decisions will help validate the work and curb doubts regarding your project.
3. Documentation is a perfect forum to showcase all of the hard work that has gone into modeling. This point cannot be emphasized enough. A large part of your success as a modeler hinges on ensuring that other departments understand the time and effort that you have put into your models. Documentation is one of many avenues that can help showcase your work.
4. In any finance/insurance context, internal/external audits are likely to occur due to the increased prevalence of model governance, model risk management, and regulatory oversight. Proper documentation will help meet these audit requirements.

Monitoring/Reporting

After your project is complete, the business will want to monitor and report the performance of the model. No model is perfect. A good monitoring program will help identify any potential issues with the final algorithm before it has a substantial impact. Additionally, monitoring will help determine when the model needs to be refreshed. As mentioned earlier in the paper, the

model is only as good as the data used to build it. As time goes on and the environment changes, the model may no longer do a good job of predicting the future. Monitoring will help identify when the modeling team should start the whole process again.

Conclusion

Applying the process above may prove very useful in your future analyses. Although the process is very time consuming, it may provide real business value to both you and your business partners. As with any project, your ability to acquire resources to embark on future projects requires an ability to clearly articulate the benefits of your analysis. If you want to embark on similar projects in the future, do not be afraid to showcase your work. The current state of technology and the availability of data make it an ideal time to start using predictive modeling in your day-to-day work.

If you have any questions, please feel free to reach out to the authors at their email addresses below:

Michael.Ewald@thehartford.com

Qiao.Wang@thehartford.com