

Quantifying Fire Risk¹

Karun Biyani

Jun Wang

Sha Yang

with the assistance of

Kristin Anderson

Jessica Smith

April 29, 2003

Abstract: For many years most insurance companies have assessed their exposure to fire losses based on a risk classification system developed by the Insurance Service Office (ISO). The loss experience has shown that the ISO system can be fallible. In this project, we propose a system to estimate the fire risk of customers more accurately, based on: (1) a list of risk factors to supplement the ISO code; and (2) a generalized linear model to analyze the influence of these factors on the incidence of fire.

¹ Work done for AAA Michigan, under the direction of Daniel Padilha, Actuarial Analyst, in partial fulfillment of the requirements of Michigan State University MTH 844, advised by Professor Peter Magyar.

Table of Contents

<i>1</i>	<i>Introduction</i>	<i>3</i>
<i>2</i>	<i>Risk Factors</i>	<i>3</i>
<i>3</i>	<i>Risk Classification Model</i>	<i>7</i>
	A. Introduction to the Generalized Linear Model.....	<i>7</i>
	B. Application of the GLM in the Project.....	<i>9</i>
<i>4</i>	<i>Analysis</i>	<i>10</i>
	A. Choosing Variables	<i>10</i>
	B. Defining Cells	<i>11</i>
	C. Selecting Interaction Terms using Backward Elimination.....	<i>13</i>
<i>5</i>	<i>Results and Interpretation</i>	<i>14</i>
<i>6</i>	<i>Conclusion and Future Work</i>	<i>20</i>
<i>7</i>	<i>References</i>	<i>20</i>
<i>8</i>	<i>Acknowledgements</i>	<i>20</i>
	<i>Appendix A. ISO Model of Fire Risk Classification</i>	<i>21</i>
	<i>Appendix B. Grouping Data into Cells</i>	<i>22</i>

1 Introduction

Risk classification is fundamental to the Insurance industry. We classify customers by their level of risk in order to avoid subsidization¹ and adverse selection² [1].

For many years most insurance companies have assessed their exposure to fire losses based on a system developed by the Insurance Service Office (ISO) (cf. Appendix A), which classifies communities into different protection classes. This system effectively distinguishes protected from unprotected communities. However, among protected communities, the ISO protection classes appear to be less effective at grouping communities in appropriate classes consistent with loss experience.

To accurately estimate the fire loss, it is important to consider separately the frequency and severity of fire. The frequency depends on the risks that can cause fire and the severity depends on the amount of insurance and the protection mechanism in place. The ISO system estimates how well protected a community is, but it is not able to identify the risk of fire occurrence associated with the community or with individual customers.

In this project, we first identify various risk factors, which may influence both frequency and severity. We then describe a generalized linear model to analyze fire risk based on those risk factors.

2 Risk Factors

The basic inputs to our fire risk model will be the risk variables associated to each customer. We use the following criteria to assess the effectiveness of risk variables:

- Separation and homogeneity of the classes
 - Each classification will have a significantly different chance of loss.
 - Each member (in a classification group) will have approximately the same chance of loss.
- Reliability
 - Information is easily obtained and not subject to manipulation.
 - Information is verifiable.
- Incentive value
 - Provides incentive to act in socially and economically positive ways.
- Social acceptability

¹ Subsidization causes people to pay more than their 'fair' share

² Undisclosed information causes people to pay less than their 'fair' share

- The Mathematically fair outcome may conflict with social goals. Some rating criteria are socially or legally unacceptable because it is beyond the insured's control.

We group the factors affecting fire risk into two types:

Risk associated with place:

- Historical experience of fire in the area,
- Number of arsons in the area,
- Water pressure,
- Fire dispatch capability,
- Fire department capability, and
- Distance from fire department.

Risk associated with home:

- Age of the home,
- Type of construction,
- Type of cooking apparatus,
- Source of heating, and
- Number of occupants and their ages.

Risk associated with place can be collected through research and survey from organizations such as the ISO. Risk associated with the home can be collected from individual customers.

In Table 1, we detail the eleven variables we consider most promising in predicting fire risk.

Table 1. Significant Risk Factors.

Variable	Significance of Variable	Method of Measurement	Reliability	Type of Variable
<i>History of Fire</i>	History of fires in an area can help in predicting the fire risk associated with that area	Statistics from Michigan State Police, State Fire Marshal;	Very reliable; fire incidents reports to State Fire Marshal are mandatory	<i>Quantitative</i> -units: number of fires per county
<i>Arson</i>	Arson is the second leading cause of fire in the United States, according to the U.S. Fire Administration	Statistics from Michigan State Police, State Fire Marshal; arson/suspicious fires by county http://www.michigan.gov/documents/2001Arson-Suspicious_54787_7.pdf	Very reliable; fire incident reports to State Fire Marshal are mandatory	<i>Quantitative</i> -units: number of arsons per county
<i>Source of Heating</i>	Heating is the third leading cause of fire in the United States, according to the U.S. Fire Administration	Required field to be filled out on application for homeowner's insurance	Reliable; applicants can easily identify source of heating	<i>Categorical</i> -oil heater -gas heater -electric heater -wood heater
<i>Cooking Apparatus</i>	Cooking is the leading cause of fire in the United States, according to the U.S. Fire Administration	Required field to be filled out on application for homeowner's insurance	Reliable; applicants can easily identify type of cooking apparatus	<i>Categorical</i> -gas -electric
<i>Age of Home</i>	Can give a good estimate of electrical distribution (i.e. safety features available at the time the home was built); also gives estimate of age of heating source	Required field to be filled out on application for homeowner's insurance	Reliable; applicants can easily access information about the age of the home	<i>Quantitative</i> -units: years

<i>Construction Type</i>	Some construction materials are more susceptible to greater amounts of damage (for example, homes built out of primarily wood may be more flamboyant than homes built out of primarily brick)	Required field to be filled out on application for homeowner's insurance	Reliable; applicants can easily observe the construction type of the home	<i>Categorical</i> -brick -wood frame -aluminum siding
<i>Distance to Nearest Fire Department</i>	Fire loss may be minimized if home is located near a fire department	Required field to be filled out on application for homeowner's insurance or use of internet (i.e. www.mapquest.com)	Very reliable; easily measurable data	<i>Quantitative</i> -units: miles
<i>Water Pressure</i>	Higher water pressure may reduce fire loss due to ability to suppress fire more quickly	Data from ISO [2]	Very reliable; easily measurable data	<i>Quantitative</i> -units: pounds per square inch
<i>Fire Department</i>	More modern and higher quality equipment and well-trained personnel may result in less fire loss	Data from ISO	Reliable; may be opinion based	<i>Categorical</i>
<i>Fire Dispatch</i>	Superior handling of fire alarms may reduce fire loss	Data from ISO	Reliable; data may be opinion based	<i>Categorical</i>
<i>Number of Occupants and Ages</i>	More people may cause increased fire risk; children under the age of 10 and senior citizens over the age of 65 are more susceptible to negligence involving fire, according to the United States Fire Administration; children playing with fire is a leading cause of fire loss, according to the Michigan State Fire Marshal	Required field to be filled out on application for homeowner's insurance	Reliable; information easily accessible to applicant	<i>Quantitative</i> -units: years

Some other risk factors that can also be considered in analyzing fire risk associated to a home are:

- School district code – There may be a correlation between the funding a school district receives and the amount of fire loss within that community.
- Market value of home – A higher market value of a home may cause greater amount of loss.
- Security system – Some security systems are equipped with the ability to detect smoke and fire and send notification to the fire department.
- Owner Occupied – An owner of a home is more likely to be more responsible than a renter.
- Smoking, alcohol habits of residents – Smoking and alcohol habits have been shown to be one of the causes that can lead to fire.
- Weather patterns (lightning, wind) – Lightning has been found to be one of the causes of fire. High wind in an area can amplify the fire as well.
- Safety features - Smoke detectors, fire extinguishers, and automatic water sprinklers are important in controlling the fire.
- Loss statistics – Previous loss experience in an area can be used to predict the occurrence of fire.

3 Risk Classification Model

In this section, we define a model to estimate fire risk based on the associated risk factors. For simplicity, we focus on the *occurrence* of fire rather than its severity.

We give a function $Y(X)$ which takes as input X , a list of risk factors for a customer, and gives the output $p = Y(X)$, a predicted yearly probability for the customer to experience a fire. Our function will involve parameters β , which are optimized using a historical database of fire occurrence.

We construct our function $Y(X)$ according to the Generalized Linear Model (GLM) [3]. As an extension of the classical linear model, GLM is used in a wide variety of statistical applications. In this section, we first introduce the generalized linear model and then describe our application of it.

A. Introduction to the Generalized Linear Model

The classical linear model works well in some problems involving a linear relationship between input variables and response variables. However, in many cases, this direct linear relationship cannot be found, either because of the distribution of the input variable or the range of the response variable, or more generally, both. Thus, the generalized linear model was developed.

In its simplest form, a linear model specifies the linear relationship between a response (or dependent) variable Y , and a set of predictor variables, the X 's, so that $X = (1, X_1, \dots, X_k)$, and:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon = X \cdot \beta + \varepsilon. \quad (1)$$

In equation (1), β_0 is the coefficient for the intercept, the β_i values are the coefficients (for the variables X_1 through X_k), and e stands for the error variability that cannot be accounted for by the predictors. The distribution of e is normal with mean zero. The optimal value of the coefficients β_i will be computed by regression from the historical data.

For many data analysis problems, estimates of the linear relationships between variables are adequate to describe the observed data, and to make reasonable predictions for new observations. However, there are many relationships that cannot adequately be summarized by a simple linear equation, for two major reasons:

(1) *Distribution of dependent variable*: The dependent variable of interest may not have a normal distribution, thus the classical linear model will not apply.

(2) *Link function*: The linear model might be inadequate to describe a particular relationship where the effect of the predictors on the dependent variable may not be linear. The generalized linear model can be used to predict responses both for response variables with discrete distributions and for response variables that are nonlinearly related to the predictors.

The relationship in the generalized linear model is assumed to be

$$g(E[Y]) = \eta,$$

where $E[.]$ denotes the expectation and g is a monotone function of

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k.$$

That is, if h is the inverse function of g , meaning $E[Y] = h(g(E[Y]))$, then

$$E[Y] = h(\eta) = h(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k). \quad (2)$$

Link Functions and distributions

Various link functions [3] can be chosen, depending on the assumed distribution of the Y variable. For example, the generalized linear model becomes the classical linear model if the link function is the identity link $g(E[Y]) = E[Y]$.

In our problem the response variable Y is *Binary*: 1 for fire and 0 for no fire. The expectation of Y is denoted by p . The commonly used link functions for the *Bernoulli* (p) distributions are:

- Logit link: $g(p) = \log(p/(1-p))$,
- Probit link: $g(p) = \text{invnorm}(p)$, where *invnorm* is the inverse of the standard normal cumulative distribution function,
- Complementary log-log link: $g(p) = \log(-(\log(1-p)))$,

- Log-log link: $g(p) = \log(\log(-p))$

We use the logit link function in our model because of its practical convenience and its simple interpretation as the logarithm of the odds ratio¹. Such a model is called the logistic regression model.

Estimation in the generalized linear model

The optimal values of the parameters (β_0 through β_k and the scale parameter) in the generalized linear model are obtained by the Maximum Likelihood (ML) method, which requires iterative computational procedures. There are many iterative methods for ML estimation in the generalized linear model, of which the Newton-Raphson and Fisher-Scoring methods are among the most efficient and widely used [4].

Maximum Likelihood Method

The maximum likelihood method is a general method of estimating parameter β of a population by values that maximize the *likelihood* (L) of a sample. The likelihood L of a sample of n observations y_1, y_2, \dots, y_n , is the joint frequency function or density function $f(y_1, y_2, \dots, y_n | \beta)$. The maximum likelihood estimation of β is the value of β that maximizes the likelihood. To maximize the likelihood L , we solve the equation $\frac{\partial L}{\partial \beta} = 0$ to get estimator β^*

such that the maximum value of L was reached at β^* . Since this equation will generally be a quite complicated nonlinear equation, we can only solve it by an iteration method. Practically solving it efficiently will involve the Newton-Raphson iteration method [5].

We apply a logistic model to our problem and try to find the appropriate function to predict the fire frequency. We denote the response variable Y , claim of fire, as a binary variable: 0 denotes no fire, 1 denotes otherwise.

B. Application of the GLM in the Project

In our project, almost all of the input data are categorical data. Categorical data are different from quantitative data, since they do not necessarily have order in them. Therefore, in this case, just assigning a number to each different type will not be the best approach, since the ordinary number will assume those categories are already ordered. Instead, we use the dummy variables to transform the j -th categorical variable with P_j categories into a P_j -dimension vector $(x_1^j, x_2^j, \dots, x_{P_j}^j)$, where $x_s^j = 1$ if the j -th variable is in its s -th category and $x_s^j = 0$ otherwise. Correspondingly, the vector β in this categorical case becomes a longer vector

$$\beta = (\beta_1^1, \beta_2^1, \dots, \beta_{P_1}^1, \beta_1^2, \beta_2^2, \dots, \beta_{P_2}^2, \dots, \beta_1^r, \beta_2^r, \dots, \beta_{P_r}^r)^T,$$

¹ Odds ratio is $\frac{p}{1-p}$, where p is the probability. We have $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$.

where each segment of $\beta_1^j, \beta_2^j, \dots, \beta_{p_j}^j$ denotes the weight of each category in the j -th variable.

We have the following equation to calculate the predicted fire frequency:

$$p = \frac{e^\eta}{1 + e^\eta} = \frac{\exp\left(\sum_{j=1}^m \sum_{k=1}^{P_j} x_{j,k} \beta_k^j\right)}{1 + \exp\left(\sum_{j=1}^m \sum_{k=1}^{P_j} x_{j,k} \beta_k^j\right)}, \quad (3)$$

where η is given by equation (2) of Section 3a.

Another advantage of this model is that one can consider *interactions* of predictors. An interaction¹ is the variation among the differences between means for different levels of one variable over different levels of the other variable. This interaction effect on response variables between different parameters, say, variable r and s , can be analyzed by adding an extra input vector of length $P_r \cdot P_s$, and the corresponding segment in estimator β .

4 Analysis

In this section, we describe the three steps in model building and apply the logistic regression to the data set. We use SAS² for the analysis of our model.

Step A. Choosing Variables

Territory is one of the important factors that affect the risk of fires. For example, there have been more occurrence of fires in Detroit than most other places in Michigan. There are several variables in the data set that reflect the risk of fires with respect to territory. Obviously, there is no need to choose all of them.

Form is another risk factor that we have chosen. It has been observed that fire is more frequent in one- or two- family dwellings than in apartments or condominiums [6]. There are 8 different Form types in the data set, which are denoted by H₁, H₃, H₄, H₅, H₆, H₇, M₁ and M₃. The distribution of the observations in these Forms is not homogeneous. There are fewer numbers of observations in type H₁ (less than 1000) and M₁ (less than 1500).

Another important factor that we should not ignore is the construction type. For example, a wood-frame home is more combustible than a brick house. There are 6 Construction types: 1-Frame, 2-Masonry Veneer, 3-Masonry, 4-Fire Resistant, 5- Aluminum Siding over, 6 - Mobile home.

¹ For example, a cholesterol reduction clinic has two diets A, B, and one exercise regime. It was found that exercise alone was effective, and diet alone was effective in reducing cholesterol levels. Also, for those patients who didn't exercise, the two diets worked equally well; those who followed diet A and exercised got the benefits of both. However, it was found that those patients who followed diet B and exercised got the benefits of both plus a bonus, an interaction effect.

² SAS (Statistical Analysis System) (SAS Institute, Inc., Cary, NC) is software used to perform statistical analysis on large sets of numerical and character data.

Most of the risk factors in the data set are categorical variables. We should be very careful about the number of variables we choose to analyze and the categories within each categorical variable. We will explain the reason in the next section.

Step B. Defining Cells

We define a cell to be a combination of several categorical variables under consideration. There is a trade off between the number of categories we want to differentiate within a categorical variable and the precision of the estimations.

Criteria for defining cells

We could use the proportion of fires in a cell to describe the probability of having fires for each observation in the cell, assuming that the observations in the cell are homogenous. The simplest method to estimate the proportion is to use the sample proportion from a sample of n observations, denoted by \hat{p} ($\hat{p} = \# \text{ fires} / \# \text{ observations}$). The accuracy of the estimation will obviously be affected by the sample size n . A sample of one fire out of 1000 observations gives the same estimation as a sample of 10 fires out of 10000 observations. But the latter estimation will be more accurate. The standard deviation of the estimation \hat{p} , denoted by

standard error, can be approximated by $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, assuming the number of observations in

the cell is large. Over the entire data set, the average probability of having at least one fire in each observation is $\hat{p} = 0.5\%$, if we want the standard error of the estimation to be within 0.1%, we will need a sample of size as large as 5000 observations; and if we want the standard error to be within 0.05%, we will need a even larger sample of size 20000 observations.

If we use too many categories of predictors, we may grid the data set into too many cells, some of which may not have enough observations in them. Imbalance of the data, i.e. data with relatively few responses of one type (relatively few fires in the data set), limits the number of categories that we can use for the model. In order to enhance the precision and at the same time incorporate most of the data into of our analysis, we group certain categories within a variable in order to have enough observations in each cell. On the other hand, no matter how we define the cells, the number of observations in the cells will not be close. There might be some large cells containing 100,000 observations. At the same time, there might be a small cell containing only a few observations. For the large cells, we can get pretty accurate estimations of the probabilities of having fires using the sample proportion of observations with at least one fire. But this simple method is not justified when applied to small cells. The standard error of the estimation will be too large.

Effect of sample size on the precision of estimations

As a preliminary analysis, we first use territory as a single predictor and estimate the probability of occurrence of fires by the sample proportion of fires observed in each territory.

In Figure 1, we plot the sample proportion of fires with the 95% confidence intervals¹ for each territory. One can observe that the 95% confidence intervals of some territories are below/above the 0.5% average probability of having a fire, which means those territories are relatively safe/risky. But there are also many territories with quite wide confidence intervals. This cast a doubt on the stability of those predictions. If the confidence intervals are too wide, many of them will overlap with each other. In other words, many of the predictions are not much different from each other. Those with shorter confidence intervals are for the territories with large number of observations, for which the above analysis is sound and stable, but such analysis cannot be applied to the territories with small number of observations.

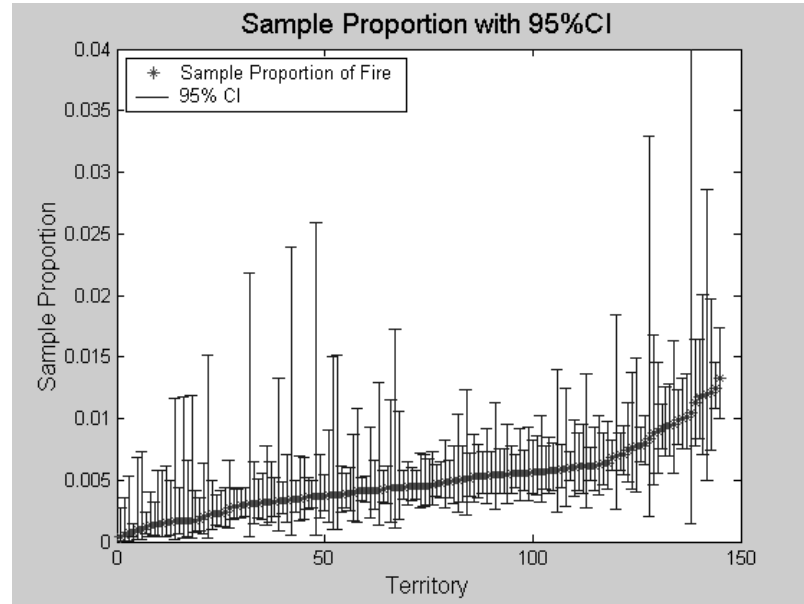


Figure 1. Sample Proportion of Fire with 95%CI

Grouping categories within each categorical variable

As mentioned earlier, we need to define the categories of the variables we are going to use in our analysis. We may also need to group certain categories within a variable in order to have enough observations in each cell.

For territory, we must differentiate risky areas from safe ones. From the map of fire claim counts², it is very obvious that Detroit has more fires than other places in Michigan. We

¹ The 95% Confidence interval for the estimators $\text{logit}(\hat{p}_i) = \frac{\hat{p}_i}{1 - \hat{p}_i}$ is defined to be $\hat{\beta} \pm 1.96(SE)$.

$SE^2 = \text{diag}\{\text{Cov}(\hat{\beta})\}$, $\text{Cov}(\hat{\beta}) = (X^T W X)^{-1} \{1 + O(n^{-1})\}$, where X is the matrix of the explanatory variables.

$W = \text{diag}\{n_i p_i (1 - p_i)\}$, where n_i is the number of observations in each cell and p_i is the predicted probabilities.

² Map provided by AAA Michigan

also noticed that there are some counties¹ (marketed in darker colors in the map) that have more fires than the rest. We introduce a variable NewT: NewT is “1” for an observation in Detroit or these riskier counties and “0” for an observation in other places.

For Form, we keep the original definition for each type, namely: H₁, H₃, H₄, H₅, H₆, H₇, M₁ and M₃. For Construction type, we group types 4, 5 and 6 together, assuming these types are very close in terms of fire risks.

The grouping of the categories of above variables is exhibited in Table 2.

Table 2. Grouping categories within variables.

	<i>Categories</i>	<i>Categorical variables</i>
Form	Form H ₁ , H ₃ , H ₄ , H ₅ , H ₆ , H ₇ , M ₁ , M ₃	unchanged
Territory	Detroit and riskier counties	NewT= 1
	Other places	NewT = 0
Construction Type	Frame	NewCon = 1
	Masonry Veneer	NewCon = 2
	Masonry	NewCon = 3
	Aluminum, Fire Resistive and Mobile home	NewCon = 4

The total number of cells (cf. Table 6 in Appendix B) based on the above design is 52. Among them 18 cells have more than 10,000 observations. The total number of observations in Form type H₁ and M₁ are not large enough to analyze. Hence, we will ignore them from our analysis. Some other criteria must be used for predicting the risk of these types.

Step C. Selecting Interaction Terms using Backward Elimination

Because our model may include interactions between the selected X_i , it may have a very large number of parameters. Our next task is to select which of these parameters are most important in influencing the response variable Y .

What is a good model? At first sight it might seem that a good model is one that fits the observed data very well. However, by including a sufficient number of parameters in our model, we can make the fit as close as we please. In fact, we can do so in a saturated model, which has as many parameters as observations. In order to do that, we reduce no complexity of the data. There are two competing goals in model selection: The model should be complex enough to fit the data well and it should also be parsimonious in parameters. A parsimonious model will give substantially better predictions than one that include unnecessary parameters. Hence, it will be robust in the sense that it will not be significantly affected by the input (possibly error).

¹ Riskier counties: OTSEGO, GRAND TRAVERS, MIDLAND, SAGINAW, KENT, GENESEE, LIVINSTON, KALAMAZOO, WASTENAW.

But, how do we know that a parameter is unnecessary? SAS has a built-in backward elimination algorithm to search for a good model. Backward elimination begins with a complex model (including all main effects and interaction effects) and sequentially removes predictors from the model. At each stage, it selects the term for which removal of the term has the least damaging effect on the model (e.g., largest p values¹). The process stops when any further deletion leads to significantly poorer fit. As a result, our model only include main effects, the interaction of Form*NewCon and NewCon*NewT.

Another well known criteria for selecting a good model is the Akaike information criterion (AIC), which is defined to be:

$$\text{AIC} = -2(\text{maximized log likelihood} - \text{number of parameters in the models}).$$

A simple model that fits the data adequately well has the advantages of parameter parsimony. Thus, AIC penalizes a model for having too many parameters. We choose a model that has the smallest AIC value. The result is consistent with the Backward Elimination procedure.

5 Results and Interpretation

In Table 3, we give the SAS output for the estimation of the parameters by the maximum likelihood method. Using the equation (3) from Section 3b, we can compute the predicted probabilities of having fires in a certain cell.

We note that in the vector for a categorical variable $(x_1^j, x_2^j, \dots, x_{p_j}^j)$, the components are not independent because the variable has to fall in one and only one category anyway. Therefore, the sum of the components equals one, i.e. $\sum_{s=1}^{p_j} x_s^j = 1$. So the estimates β_s^j associated with each x_s^j are not independent either. We may put an additional constraint to uniquely determine β_s^j . In statistics it is common practice to set the estimate of the last category to be zero, i.e. $\beta_{p_j}^j = 0$. Hence, the estimates of type M₁, NewT = 1 and NewCon = 4 have been set to zero, respectively.

Using our model,

$$\eta = \beta_{\text{Form}} + \beta_{\text{NewT}} + \beta_{\text{NewCon}} + \beta_{\text{Form*NewT}} + \beta_{\text{NewCon*NewT}},$$

where β_{Form} , β_{NewT} , etc. are determined from the Table 3.

The predicted probability for each cell is $\hat{p} = \frac{e^\eta}{1 + e^\eta}$.

For example, the probability of having fires for an observation in the cell with NewForm=H₃, NewCon=1, NewT=0 is:

¹ Null hypothesis $H_0: \beta = 0$, where β is the coefficient of one of the predictors in our model. “ $\beta = 0$ ” means that the response variable and the predictors are independent. p value is defined to be the probability of the observed event (response) to happen if the null hypothesis is correct. Therefore, we will accept H_0 if the p value is large and reject it if p value is small.

$$\hat{p} = \frac{e^\eta}{1+e^\eta} = \frac{e^{-5.1607}}{1+e^{-5.1607}} \approx 0.0057.$$

In Table 4, we give the predicted probabilities for each of 42 cells that we defined based on the model. (Among the 52 cells (cf. Appendix B), we ignore Form H₁ and M₃ that don't have enough observations to estimate).

Table 3. Analysis of Maximum Likelihood Estimators.

<i>Predictors</i>	<i>Estimators</i>					
	<i>Values of Predictors</i>	<i>DF</i>	<i>(β)</i>	<i>Standard Error</i>	<i>Chi-Square</i>	<i>Pr > ChiSq</i>
Intercept		1	-5.5514	0.2587	460.4844	<.0001
NewT	0	1	-0.2651	0.3168	0.7003	0.4027
Form	H3	1	0.4013	0.2736	2.1502	0.1425
Form	H4	1	-0.6776	0.2931	5.3434	0.0208
Form	H5	1	-0.1033	0.2779	0.1383	0.7099
Form	H6	1	-0.8626	0.3458	6.223	0.0126
Form	H7	1	0.4947	0.2874	2.9629	0.0852
Form*NewT	H3 0	1	0.0501	0.3351	0.0224	0.8811
Form*NewT	H4 0	1	-0.4311	0.3647	1.397	0.2372
Form*NewT	H5 0	1	0.4467	0.3391	1.7358	0.1877
Form*NewT	H6 0	1	0.2458	0.409	0.3612	0.5478
Form*NewT	H7 0	1	-0.2559	0.3926	0.4247	0.5146
NewCon	1	1	0.3261	0.0964	11.4463	0.0007
NewCon	2	1	0.3962	0.1013	15.3122	<.0001
NewCon	3	1	-0.0326	0.205	0.0252	0.8738
NewCon*NewT	1 0	1	-0.1217	0.1183	1.0588	0.3035
NewCon*NewT	2 0	1	0.1569	0.122	1.6537	0.1985
NewCon*NewT	3 0	1	-0.1987	0.2445	0.6608	0.4163

Table 4. Predicted probabilities for the 42 cells.

<i>Cell</i>	<i>Form</i>	<i>NewCon</i>	<i>NewT</i>	<i>Fire</i>	<i>Non Fire</i>	<i>Total</i>	<i>Fire/Total</i>	<i>Predicted Risk</i>	<i>95% CI</i>	
									<i>lower</i>	<i>upper</i>
1	H4	3	0	5	2908	2913	0.001716444	0.000779165	0.000575529	0.001055
2	H4	4	0	5	5944	5949	0.000840477	0.000981739	0.000780987	0.001234
3	H4	1	0	32	22086	22118	0.001446785	0.001204075	0.000980866	0.001478
4	H6	3	0	0	1467	1467	0	0.001273694	0.000936353	0.001732
5	H6	3	1	0	292	292	0	0.001583478	0.00091105	0.002751
6	H6	4	0	5	3643	3648	0.001370614	0.001604634	0.001269577	0.002028

<i>Cell</i>	<i>Form</i>	<i>NewCon</i>	<i>NewT</i>	<i>Fire</i>	<i>Non Fire</i>	<i>Total</i>	<i>Fire/Total</i>	<i>Predicted Risk</i>	<i>95% CI</i>	
									<i>lower</i>	<i>upper</i>
7	H6	4	1	0	1168	1168	0	0.001635801	0.001043995	0.002562
8	H4	2	0	59	38789	38848	0.00151874	0.001705674	0.001400716	0.002077
9	H4	3	1	4	1306	1310	0.003053435	0.00190457	0.001256159	0.002887
10	H4	4	1	3	2808	2811	0.001067236	0.001967482	0.001502307	0.002576
11	H6	1	0	15	9291	9306	0.001611863	0.001967759	0.001590684	0.002434
12	H6	1	1	12	4920	4932	0.00243309	0.002265037	0.00148533	0.003453
13	H6	2	1	10	3471	3481	0.002872738	0.002429143	0.001588027	0.003714
14	H4	1	1	40	14328	14368	0.002783964	0.002723957	0.002183954	0.003397
15	H6	2	0	75	24674	24749	0.003030425	0.002786611	0.00227663	0.00341
16	H4	2	1	40	13597	13637	0.002933196	0.002921215	0.002341188	0.003644
17	M3	4	0	30	10074	10104	0.002969121	0.002969121	0.002076719	0.004243
18	H7	3	0	2	206	208	0.009615385	0.0029914	0.001942464	0.004604
19	H5	3	0	30	8542	8572	0.003499767	0.003320118	0.002617935	0.00421
20	H5	3	1	4	2160	2164	0.001848429	0.003377239	0.002312848	0.004929
21	H5	4	1	35	9640	9675	0.003617571	0.003488626	0.002861824	0.004252
22	H3	3	0	34	10184	10218	0.003327461	0.003697358	0.002921786	0.004678
23	H7	4	0	3	787	790	0.003797468	0.003766961	0.002572228	0.005514
24	M3	4	1	15	3864	3879	0.003866976	0.003866976	0.002332569	0.006404
25	H5	4	0	109	24941	25050	0.004351297	0.004180547	0.003676079	0.004754
26	H7	1	0	20	4333	4353	0.004594533	0.004617156	0.003207202	0.006643
27	H3	4	0	155	33702	33857	0.004578078	0.004655095	0.004117175	0.005263
28	H5	1	1	158	28898	29056	0.005437775	0.004827133	0.004224394	0.005515
29	H5	1	0	286	59648	59934	0.004771916	0.005123606	0.004705412	0.005579
30	H5	2	1	65	15523	15588	0.004169874	0.005175905	0.004454795	0.006013
31	H3	3	1	20	2936	2956	0.0067659	0.005581457	0.003867204	0.008049
32	H3	1	0	587	98961	99548	0.005896653	0.005704591	0.005309935	0.006128
33	H3	4	1	79	14326	14405	0.005484207	0.005765123	0.004844983	0.006859
34	H7	3	1	1	364	365	0.002739726	0.006124674	0.004083077	0.009178
35	H7	4	1	20	1997	2017	0.009915716	0.006326102	0.004956676	0.008071
36	H7	2	0	4	808	812	0.004926108	0.006531295	0.004514337	0.009441
37	H5	2	0	545	72561	73106	0.007454928	0.007246177	0.006735103	0.007796
38	H3	1	1	278	36190	36468	0.007623122	0.007970085	0.007203585	0.008817
39	H3	2	0	402	51016	51418	0.007818274	0.008065896	0.007459085	0.008722
40	H3	2	1	188	20271	20459	0.00918911	0.008543993	0.00757214	0.009639
41	H7	1	1	34	4630	4664	0.00728988	0.008743733	0.007164189	0.010668
42	H7	2	1	59	6153	6212	0.009497746	0.009372823	0.007725577	0.011367

To test the goodness of fit, SAS provides the Hosmer and Lemeshow Test¹ with p value = 0.7848, which indicates a goodness of fit. Table 5 is a list of partition for the Hosmer and Lemeshow test. One can find that the observed occurrence of fire is very close to the expected occurrence of fire based on our model.

Table 5. Partition for the Hosmer and Lemeshow Test.

<i>Group</i>	<i>Total</i>	<i>Fire</i>		<i>No Fire</i>	
		<i>Observed</i>	<i>Expected</i>	<i>Observed</i>	<i>Expected</i>
1	70866	107	108.54	70759	70757.46
2	71866	156	154.47	71710	71711.53
3	49794	151	163.58	49643	49630.42
4	45975	213	198.11	45762	45776.89
5	272453	1540	1525.76	270913	270927.2
6	133326	957	970.17	132369	132355.8
7	34779	348	347.88	34431	34431.12

A residual plot is one of the most frequently used graphical devices for model diagnosis. Residuals are the differences between the predictions and the observations. Since the true probability of having fires in each cell is unknown, we simply compare our predication with the sample proportion of fires in each cell. Figure 2 gives the residual plot against the observed probabilities, which are simply the proportion fires in each cell. A randomly distributed small-scale residual is desired because it indicates that the model has captured all of the systematical effects. From the plot, one can observe that most of the residuals are very small and there is no pattern in the distribution, indicating randomness. But one can also observe there are a few relatively large residuals, like the one on the top (residual equals 0.0066), we track and find it is from the cell H₇(Form), 3(NewCon), 2(NewT), which only has only 206 observations. Because of such a small sample size (206), the observed sample proportion may not be a good indication of the risk associate with this cell. Hence, the residual in such a small cell does not indicate a poor fit.

If we removed the residuals from these smaller cells, one could expect the residuals to reduce. Hence, in the Figure 3, we plot the residuals of only the larger cells (Cells with more than 10000 observations). Indeed, the residual reduced dramatically, to be less than 0.0007.

Figure 4 is the plot of the proportion of fires in each cell against estimated probabilities. Figure 5 is a similar plot only for the larger cells.

¹ The subjects are divided into approximately ten or less groups of roughly the same size based on the percentiles of the estimated probabilities. The discrepancies between the observed and expected number of observations in these groups are summarized by the Pearson chi-square statistic, which is then compared to a chi-square distribution with t degrees of freedom, where t is the number of groups minus n . By default, $n=2$. A small p -value suggests that the fitted model is not an adequate model. p -value range from 0 to 1.

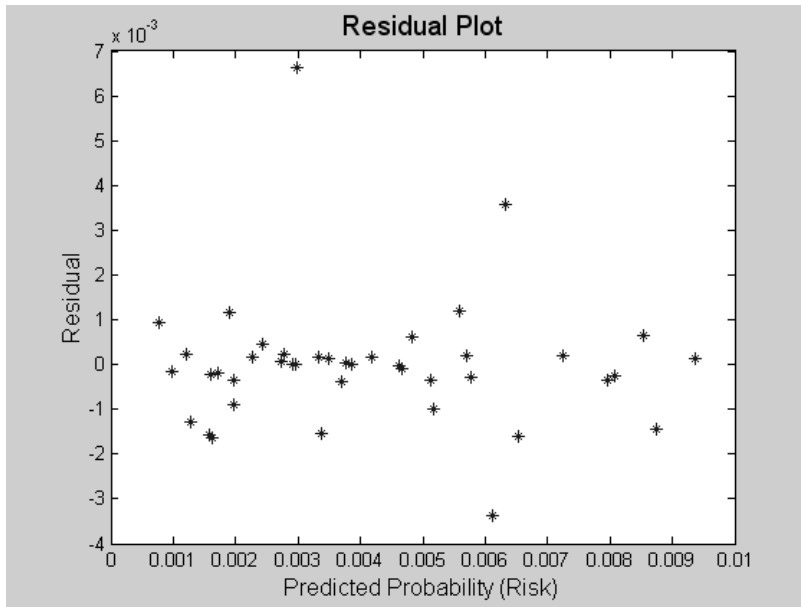


Figure 2. The Residual plot of the estimated probabilities.

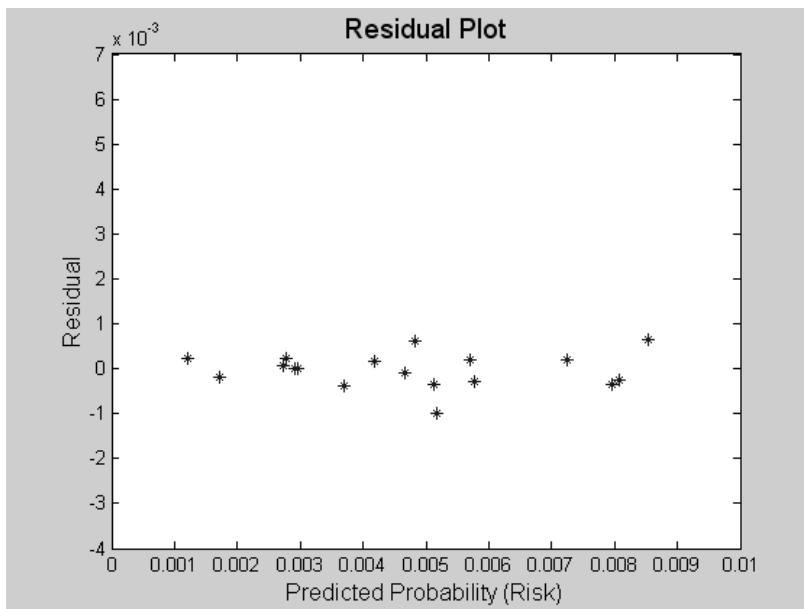


Figure 3. The Residual plot of the Estimated Probability in large cells.

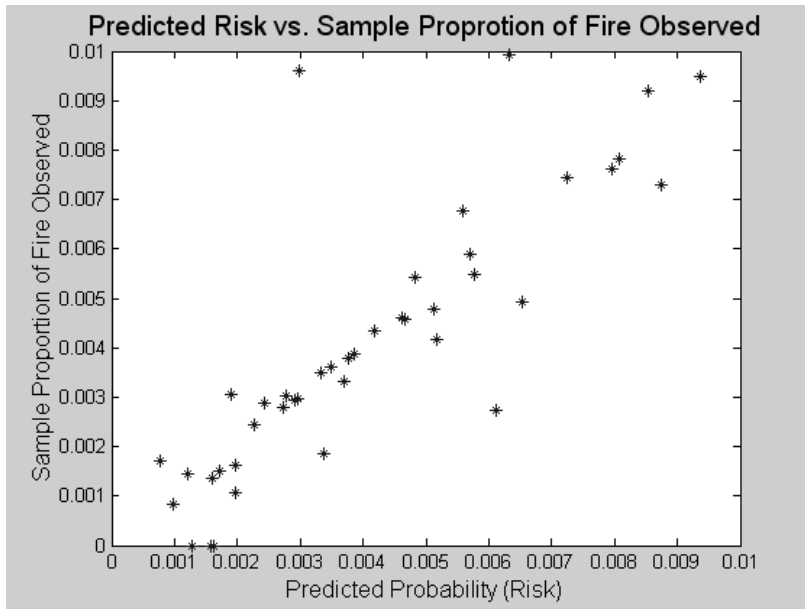


Figure 4. Estimated Probability vs. Proportion of fires in each cell.

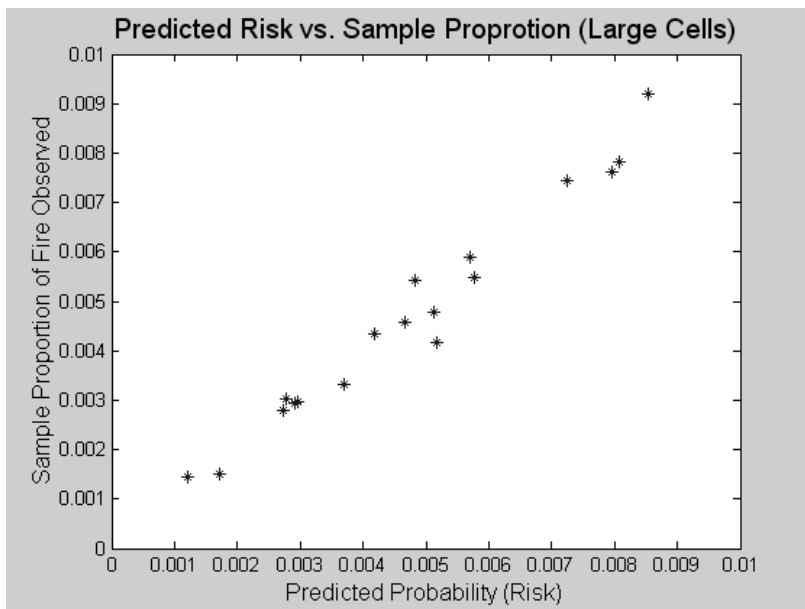


Figure 5. Estimated Probability vs. Proportion of fires in larger cells.

6 Conclusion and Future Work

In this project, we quantify the fire risk based on several risk factors. By grouping the data having the same combination of categorical risk predictors, we modeled the fire risk using generalized linear models. We found a way (cf. equation (3) in Section 3) to predict the fire risk of all combinations of predictors. The residual plots (cf. Section 4) suggest that our model does predict the fire risk accurately.

There may never be enough data to make accurate estimations for all types of customers. The significance of this model is that it can give better estimates even for those types of customers without an adequate number of historical observations.

In our current analysis, we predict only the frequency of fire. However, given more data related to amount of fire loss, such as *amount of insurance* for each policy, we can apply GLM to predict the amount of fire loss and not just the occurrence of fire.

7 References

- [1] <http://www.prenhall.com/financecenter/html/downloads/dorfman/Ch02v4.ppt>
- [2] Insurance Services Office – <http://www.iso.com/>
- [3] J.A. Nelder and P. McCullagh, 1989. *Generalized Linear Models*, 2nd edition. CRC Press.
- [4] A. J. Dobson. 1990. *An introduction to generalized linear model*. Chapman & Hall.
- [5] Alan Agresti, 2002. *Categorical Data Analysis*. 2nd edition. John Wiley & Sons.
- [6] Marty Ahrens. *The U.S fire problem overview and report*. June 2001. NFPA.

8 Acknowledgements

We would like to express our sincere gratitude to Daniel Padilha for his guidance and suggestions. We would also like to thank our faculty manager, Dr. Peter Magyar, for his continuous support throughout this project and Dr. Stapleton James for his suggestions in building the model for this project. Finally, we thank Dr. Charles MacCluer for his suggestions with the report.

Appendix A. ISO Model of Fire Risk Classification

The Insurance Services Office (ISO) [2] is the premiere source of information products, and services related to property and liability risk. The ISO provides statistical, actuarial, underwriting, and claims information and analyses; consulting and technical services; policy language; information about specific locations; fraud-identification tools; and data processing. For classifying fire risk in case of home insurance, the ISO allocates protection classes to each community. The insurance companies use these classes when determining the premiums.

Factors used in Classification

Many factors contribute to the frequency and severity of fire losses in different regions of the state. One of the major players in fire losses is the protection each community offers for the residents. Insurance companies need reliable information about fire protection services in each community within Michigan in order to help calculate accurate premiums. The Insurance Service Office (ISO) collects the necessary information and provides a classification system for a community's public fire protection. When a community is evaluated, the ISO looks at three main aspects: (1) Fire dispatch, (2) Fire department, and (3) Water supply.

Fire dispatch refers to the receiving and handling of fire alarms. This aspect accounts for 10% of the ISO rating and includes a review of the community's facilities, as well as the support for handling and dispatching fire alarms. A review of the fire department accounts for 50% of the total rating. The main focuses of this review are the first-alarm response and immediate reaction to minimize potential loss. Within these focus groups, the ISO concentrates on engine companies, ladder or service companies, distribution of fire stations and fire companies, equipment carried on the fire trucks, pumping capacity, reserve equipment, department staff, and training. The final 40% of the rating is derived from an analysis of the water supply. The analysis consists of an evaluation of the size, type, installation, inspection, and condition of fire hydrants. In addition, an inadequate water supply may limit the capability of the fire department to contain fires or an insufficient fire department may not be able to effectively utilize a copious water supply. In these cases, the ISO makes an adjustment to the total score, known as divergence, to reflect the limiting effect.

Protection Class

After these inspections are conducted and analyzed, the ISO allocates a protection class code (PC) to each city, village, and township. These codes range from one to ten, where a code of one is generally considered to have the optimum public protection and a code of ten is typically categorized as the least protected community, indicating that less than the minimum requirements have been met. However, communities with over 250,000 residents are statistically rated. These communities are assigned PC codes based on fire loss experience, without regard to engineering surveys.

The ISO provides insurance companies with accurate information regarding each community's fire protection. This facilitates the insurance companies with the evaluation and assignment of premiums.

Appendix B. Grouping Data into Cells

Based on our definition of cell (cf. Section 4), we divide the data sets in 52 cells, which are shown in Table 6. The distribution of observations among the cells is non-homogeneous. Among them 18 are larger cells (containing more than 10000 observations).

Table 6. Distribution of observations and fires in cells.

<i>Cell</i>	<i>Form</i>	<i>New Con</i>	<i>NewT</i>	<i>Fire</i>	<i>No fire</i>	<i>Total</i>	<i>Fire Proportion</i>
1	H1	1	0	1	539	540	0.001852
2	H1	1	1	0	67	67	0
3	H1	2	0	0	74	74	0
4	H1	2	1	1	97	98	0.010204
5	H1	3	0	1	58	59	0.016949
6	H1	3	1	0	8	8	0
7	H1	4	0	0	41	41	0
8	H1	4	1	0	15	15	0
9	H3	1	0	587	98961	99548	0.005897
10	H3	1	1	278	36190	36468	0.007623
11	H3	2	0	402	51016	51418	0.007818
12	H3	2	1	188	20271	20459	0.009189
13	H3	3	0	34	10184	10218	0.003327
14	H3	3	1	20	2936	2956	0.006766
15	H3	4	0	155	33702	33857	0.004578
16	H3	4	1	79	14326	14405	0.005484
17	H4	1	0	32	22086	22118	0.001447
18	H4	1	1	40	14328	14368	0.002784
19	H4	2	0	59	38789	38848	0.001519
20	H4	2	1	40	13597	13637	0.002933
21	H4	3	0	5	2908	2913	0.001716
22	H4	3	1	4	1306	1310	0.003053
23	H4	4	0	5	5944	5949	0.00084
24	H4	4	1	3	2808	2811	0.001067
25	H5	1	0	286	59648	59934	0.004772
26	H5	1	1	158	28898	29056	0.005438
27	H5	2	0	545	72561	73106	0.007455
28	H5	2	1	65	15523	15588	0.00417

29	H5	3	0	30	8542	8572	0.0035
30	H5	3	1	4	2160	2164	0.001848
31	H5	4	0	109	24941	25050	0.004351
32	H5	4	1	35	9640	9675	0.003618
33	H6	1	0	15	9291	9306	0.001612
34	H6	1	1	12	4920	4932	0.002433
35	H6	2	0	75	24674	24749	0.00303
36	H6	2	1	10	3471	3481	0.002873
37	H6	3	0	0	1467	1467	0
38	H6	3	1	0	292	292	0
39	H6	4	0	5	3643	3648	0.001371
40	H6	4	1	0	1168	1168	0
41	H7	1	0	20	4333	4353	0.004595
42	H7	1	1	34	4630	4664	0.00729
43	H7	2	0	4	808	812	0.004926
44	H7	2	1	59	6153	6212	0.009498
45	H7	3	0	2	206	208	0.009615
46	H7	3	1	1	364	365	0.00274
47	H7	4	0	3	787	790	0.003797
48	H7	4	1	20	1997	2017	0.009915716
49	M1	4	0	1	1117	1118	0.000894454
50	M1	4	1	0	167	167	0
51	M3	4	0	30	10074	10104	0.002969121
52	M3	4	1	15	3864	3879	0.003866976