Applying Graphical Models to Automobile Insurance Data

By Farrokh Guiahi

Abstract

Analysis of insurance data provides input for making decisions regarding underwriting, pricing of insurance products, claims as well as profitability analysis. In this paper, we consider graphical modelling as a vehicle to reveal dependency structure of categorical variables used in the Australian Automobile data. The methodology developed here may supplement the traditional approach to ratemaking.

Topics considered are: the description of the automobile data set; preprocessing of the variables; visualization tools suitable for contingency tables; classical test of independence; Log-Linear models; concept of conditional independence; and graphical modelling as a vehicle to explore the dependency structure among categorical variables; review of frequency rates by rating class.

Keywords:

Categorical variables, visualization of categorical variables, mosaic plots, chi-square test, Log-Linear models, conditional independence, cliques, graphical modelling, frequency rates.

1. Introduction

In a data-driven decision making environment, analysis of insurance data provides input for making decisions regarding underwriting, pricing of insurance products, claims and profitability. The focus of this paper is to study the dependency structure of categorical variables pertaining to the Australian automobile insurance data, and explore potential applications to determination of frequency rates by rating classes.

Insurers gather information about their policy holders at the time of writing insurance policies. The data collected by an insurance carrier depends on the line of business offered, business expediency, and legal constraints. The rating variables are used to price insurance products based on the insured's characteristics, and are helpful with regard to underwriting selection. Rating variables are of **mixed** measurement types.

Our focus here is only with respect to the categorical variables used in the Australian Automobile insurance data.

The analysis of insurance data is a multi-facet endeavor. The goals of statistical data analysis are broadly of two types: **understanding** and **prediction**. Understanding encompasses summarization as well as inference.

Tasks associated with machine leaning, learning from data, are categorized as **supervised learning** and **unsupervised learning**. **Predictive modeling** is an example of supervised learning, where features are used to predict the value of a target variable.

Unsupervised learning is mainly concerned with finding relationships between features, or grouping of instances as to reveal hidden underlying structure of data with no designated target variable involved.

The goal of the analysis in this paper is primarily with regard to understanding and an exercise in unsupervised learning.

We consider Exploratory Data Analysis (EDA) tools suitable for categorical variables. Inferential procedures such as tests of independence and fitting Log-Linear models to data are discussed. Furthermore, we discuss some limitations of these tools and procedures. We now proceed with briefly outlining the contents of the other sections. Description of the data used, exploratory data analysis as well as preprocessing of the data are covered in section 2. In section 3, we consider two-way contingency tables, Pearson chi-square test for assessing the strength of association between two categorical variables. Mosaic plots, as a visualization tool, to present two-way contingency tables, and we discuss some limitations of analysis based solely on two variables. Section 4, considers the analysis of three categorical variables. In particular, we extend the description of mosaic plots to that of three variables, introduce Log-linear models, the concept of conditional independence, and graphical modeling. Considerations of more than three categorical variables and model selection have been relegated to section 5. The use of graphical modeling for exploring dependency structure of categorical variables, potential applications to determining frequency rates, and overfitting are also discussed in section 5. Summary and concluding remarks are stated in section 6.

An attempt has been made to blend theory with the necessary statistical computations as to make the paper useful to practicing actuaries.

2. **Data, and Data Preprocessing**

Insurance data are generally propriety information of the insurance companies and are not publicly available. The data used in this paper is available from the following Web site and has been referenced in the book co-authored by de Jong and Heller (2008): http://www.businessandeconomics.mq.edu.au/our_departments/Applied_Finance_and_Actuarial_Studies/research/books/GLMsforInsuranceData/data_sets.

The data consisted of one year of Australian vehicle insurance policies taken out in 2004 or 2005. The name of data set is car.csv and there is a brief description of the variables considered in the file named vehicle.txt which has been reproduced here in the Appendix A.1. There were 67,856 observations, and for each record the values of ten attributes were given.

The software R has been used to perform computations and display graphics. R is open source software useful for doing statistical analysis and data visualization. The base package of R has many useful standard functions, and furthermore there are over four thousand supplemental packages which enhance the capabilities of R. Information about R and its associated packages can be obtained from http://cran.r-project.org/. Some of the R functions used in this paper are referenced in Appendix A.2 and may be of interest to practicing actuaries.

A quick “feel” of the data can be obtained from Exhibit 2.1 and Exhibit 2.2 below.

Exhibit 2.1: The Structure of Automobile data set

```
'data.frame':   67856 obs. of 10 variables:
 $ veh_value: num  1.06 1.03 3.26 4.14 0.72 2.01 1.6 1.47 0.52 0.38 ...
 $ exposure : num  0.304 0.649 0.569 0.318 0.649 ...
 $ clm      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ numclaims: int  0 0 0 0 0 0 0 0 0 0 ...
 $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
 $ veh_body : Factor w/ 13 levels "BUS","CONVT",..: 4 4 13 11 4 5 8 4 4 4 ...
 $ veh_age  : int  3 2 2 2 4 3 3 2 4 4 ...
 $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 2 1 1 ...
 $ area     : Factor w/ 6 levels "A","B","C","D",..: 3 1 5 4 3 3 1 2 1 2 ...
 $ agecat   : int  2 4 2 2 2 4 4 6 3 4 ...
```

Exhibit 2.1 provides information about the name, type of measurement, and sample values for each variable in the automobile data set. For instance, Vehicle Value (veh_value} is a continuous (num) variable

with sample values in unit of 10,000 while Vehicle Body Type (veh_body) is a nominal attribute.

Exhibit 2.2: Five Summary Statistics for the Automobile Data

```
   veh_value          exposure             clm             numclaims         claimcst0
 Min.   : 0.000   Min.   :0.002738   Min.   :0.00000   Min.   :0.00000   Min.   :    0.0
 1st Qu.: 1.010   1st Qu.:0.219028   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:    0.0
 Median : 1.500   Median :0.446270   Median :0.00000   Median :0.00000   Median :    0.0
 Mean   : 1.777   Mean   :0.468651   Mean   :0.06814   Mean   :0.07276   Mean   :  137.3
 3rd Qu.: 2.150   3rd Qu.:0.709103   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:    0.0
 Max.   :34.560   Max.   :0.999316   Max.   :1.00000   Max.   :4.00000   Max.   :55922.1
    veh_body        veh_age      gender    area          agecat
 SEDAN  :22233   Min.   :1.000   F:38603   A:16312   Min.   :1.000
 HBACK  :18915   1st Qu.:2.000   M:29253   B:13341   1st Qu.:2.000
 STNWG  :16261   Median :3.000             C:20540   Median :3.000
 UTE    : 4586   Mean   :2.674             D: 8173   Mean   :3.485
 TRUCK  : 1750   3rd Qu.:4.000             E: 5912   3rd Qu.:5.000
 HDTOP  : 1579   Max.   :4.000             F: 3578   Max.   :6.000
 (Other): 2532
```

Exhibit 2.2 provides crude information with regard to whether a variable is symmetrically distributed, appearance of outliers, and if there are any inconsistent observed values.

The analysis in this paper is primarily concerned with the categorical--nominal or ordinal--attributes. The preprocessing of the data involved: a) change of names of the variables used for easier references, b) transformations of two of the variables, and c) selecting seven out of the ten variables for this study. Additional details are provided in Appendix A.3.

The variable “Value” was originally recorded on a numeric scale. Value is used as a proxy for “size” and or “power” of a vehicle and serves as an underwriting variable. Since our primary interest is with categorical variables, it was decided to include Value in our analysis as an ordinal rather than numeric variable. The process of transforming a numeric variable to an ordinal variable is referred to as binning or feature-discretization, see Kantardzic (2011). There are two commonly methods used for binning. One method uses equal frequency and the other one uses equal length. Here, we chose the equal frequency approach. The 25% quartile (1.01), 50% quartile (1.50) and 75% quartile (2.15) of the Value were selected as cutoff points for binning.

The other transformed variable was the “Body”. The categorical variable Body had originally thirteen levels. Some of the levels had relatively low frequencies, see Table 2.1 below. These low frequency classes can affect the results of some statistical procedures used to

fitting a model to the data. Reducing the number of levels, collapsing, may be achieved subjectively based upon an expert's opinion, or it may be determined objectively based on certain criterion. I decided to consider the variable Claim in conjunction with Body to do reduce the number of levels to be used, see Table 2.1.

Table 2.1: Body and Claim
Frequency and Relative Frequency Summary

	Body\Claim	No Claim	Claim	No Claim	Claim
1	BUS	39	9	0.00062	0.00195
2	CONVT	78	3	0.00123	0.00065
3	COUPE	712	68	0.01126	0.01471
4	HBACK	17,651	1,264	0.27915	0.27336
5	HDTOP	1,449	130	0.02292	0.02811
6	MCARA	113	14	0.00179	0.00303
7	MIBUS	674	43	0.01066	0.00930
8	PANVN	690	62	0.01091	0.01341
9	RDSTR	25	2	0.00040	0.00043
10	SEDAN	20,757	1,476	0.32827	0.31920
11	STNWG	15,088	1,173	0.23861	0.25368
12	TRUCK	1,630	120	0.02578	0.02595
13	UTE	4,326	260	0.06842	0.05623
	Total	63,232	4,624	1.00000	1.00000

For any two levels of the Body type, we can measure their "closeness" by using the Euclidean distance based on the relative frequencies. For instance, to see how close is "BUS" to "CONVT", the Euclidean distance is computed as follows:

$$d(BUS, CONVT) = \sqrt{(0.00062 - .00123)^2 + (0.00195 - 0.00065)^2} = 0.00144$$

The distance between any two distinct levels of the Body has been computed and organized as a distance matrix in Exhibit 2.3.

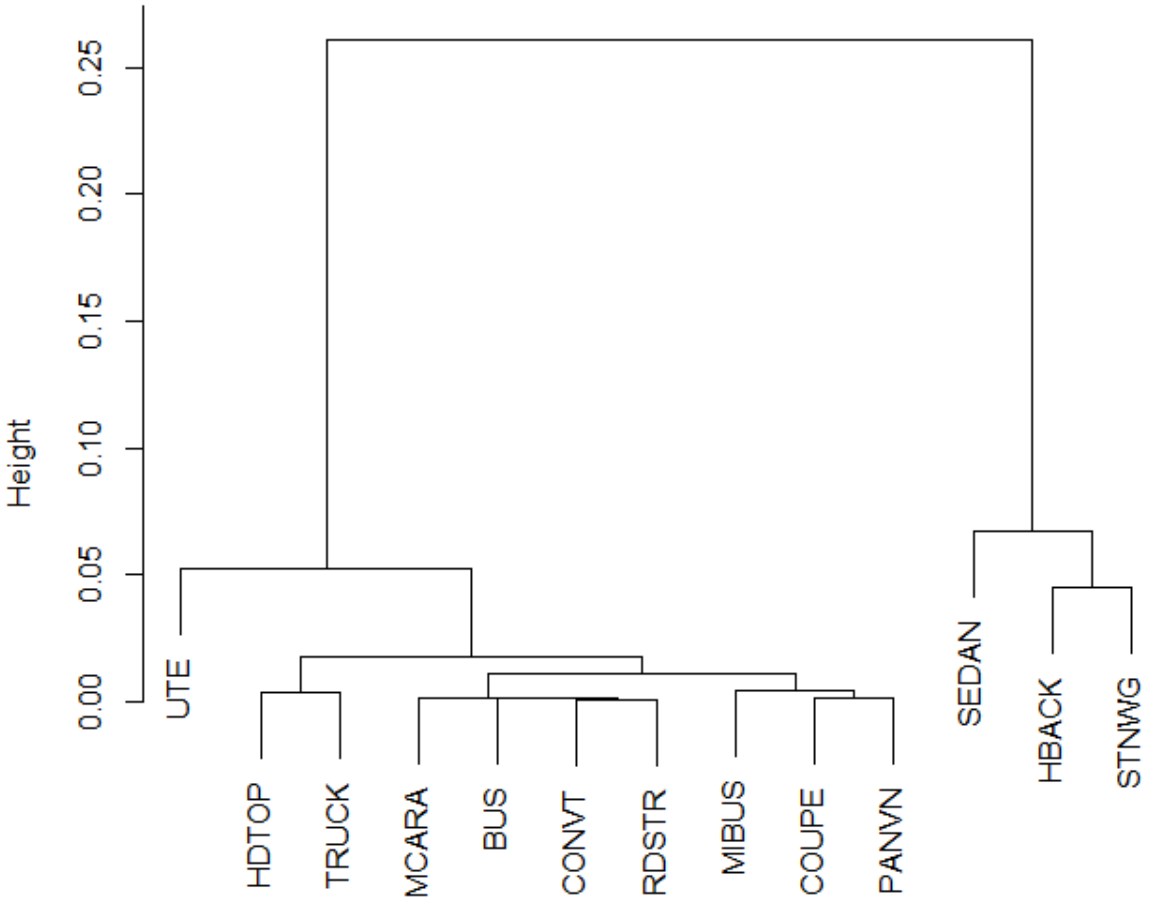
Exhibit 2.3: Distance Matrix for the Body Type

	BUS	CONVT	COUPE	HBACK	HDTOP	MCARA	MIBUS	PANVN	RDSTR	SEDAN	STNWG	TRUCK
CONVT	0.00144											
COUPE	0.01662	0.01727										
HBACK	0.38890	0.38937	0.37238									
HDTOP	0.03438	0.03499	0.01777	0.35468								

MCARA	0.00159	0.00244	0.01504	0.38731	0.03280							
MIBUS	0.01245	0.01279	0.00544	0.37658	0.02245	0.01086						
PANVN	0.01541	0.01602	0.00134	0.37353	0.01898	0.01382	0.00412					
RDSTR	0.00153	0.00087	0.01794	0.39011	0.03569	0.00294	0.01356	0.01670				
SEDAN	0.45608	0.45654	0.43956	0.06719	0.42187	0.45449	0.44375	0.44071	0.45729			
STNWG	0.34643	0.34695	0.32984	0.04506	0.31210	0.34484	0.33419	0.33102	0.34768	0.11105		
TRUCK	0.03478	0.03525	0.01836	0.35413	0.00359	0.03318	0.02249	0.01945	0.03599	0.42130	0.31170	
UTE	0.08685	0.08719	0.07065	0.30258	0.05348	0.08526	0.07442	0.07169	0.08798	0.36970	0.26068	0.05229

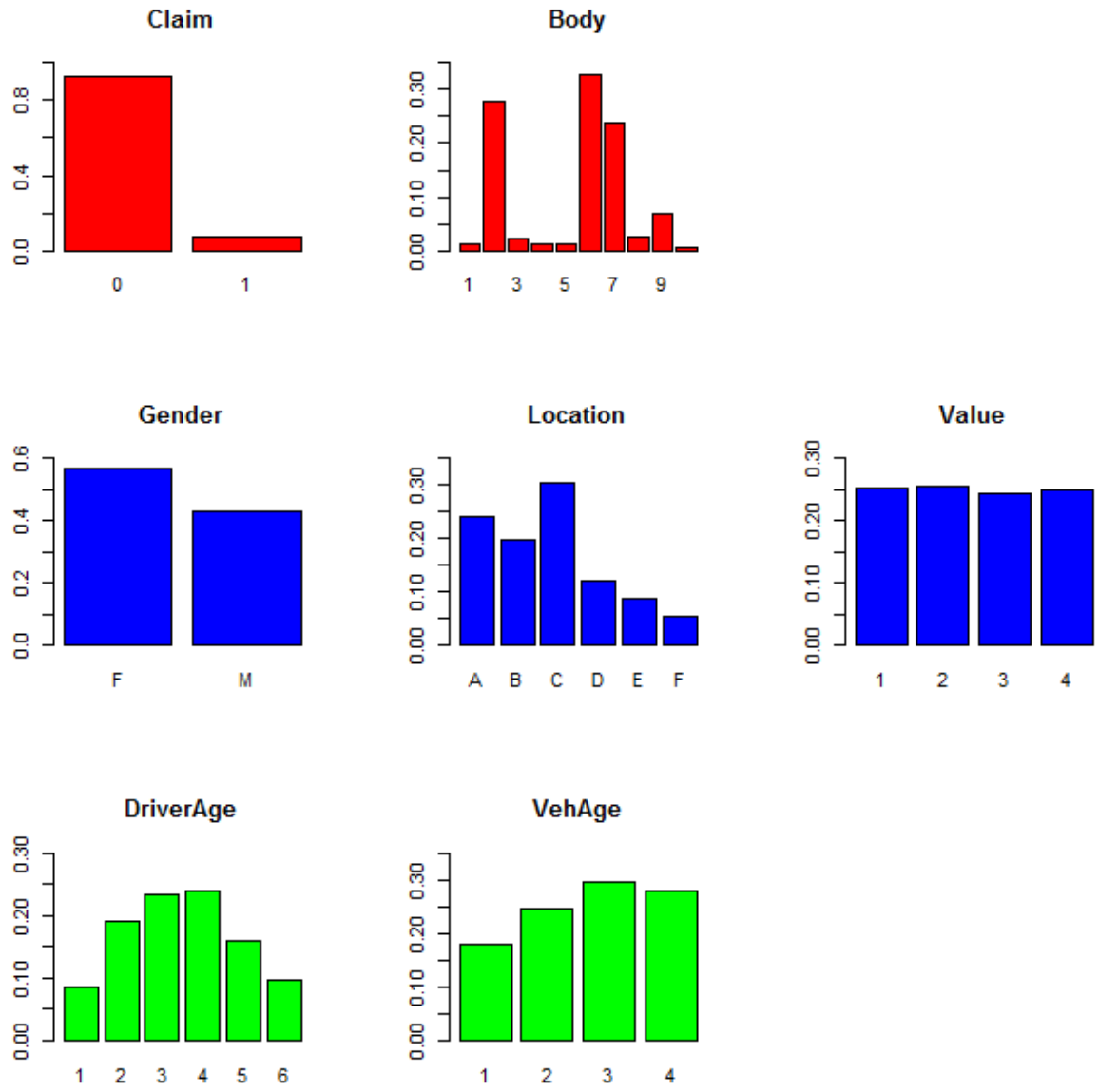
The matrix of distances is used as an input to a hierarchical clustering algorithm for the purpose of grouping levels of the Body. For an explanation of the hierarchical clustering method refer to Johnson and Wichern (2007). The output of the hierarchical analysis is a dendrogram (an inverted tree), see Exhibit 2.4 below. The lower section of the dendrogram, suggests combining the low frequency levels BUS, CONVT, MCARA and RDSTR into a single class labeled “Other”. Hence, number of levels of Body was reduced from thirteen to ten. Furthermore, the label of the levels was changed from character to numeric type in order to facilitate graphical presentation.

Exhibit 2.4: Dendrogram



Bar charts are commonly used to show the frequency distributions of categorical variables. In Exhibit 2.5, we display the Bar charts for the attributes studied here, based on their **relative frequencies.**

Exhibit 2.5: Bar Charts for Relative Frequency of Attributes Studied



Now, we may proceed with the statistical analysis phase of the paper.

3. **Analysis of Two Categorical Variables**

In this section, we discuss exploratory tools as well as statistical test of independence for two categorical variables. The information about two categorical variables is summarized as a two-way contingency table obtained by cross-tabulating the data. The contingency table for

Value and Vehicle Age is given in Exhibit 3.1. A mosaic plot, explained below, the top-left panel of Exhibit 3.2 provides a graphical display corresponding to the contingency table in Exhibit 3.1.

Exhibit 3.1

VehAge	1	2	3	4
Value				
1	3	202	4964	11891
2	2175	3897	7112	4124
3	3365	6883	3734	2568
4	6714	5605	4254	365

The association between two categorical variables is determined by conducting a Pearson chi-square test of independence.

Let us proceed with introducing the necessary notations and terms needed to perform the Pearson chi-square test.
For two categorical variables A and B, having domains: $dom(A) = \{1,2,\ldots,J\}$ and $dom(B) = \{1,2,\ldots,K\}$, the subset of the data with $A = j \text{ and } B = k$ is labelled as the cell $(j,k) \in dom(A) \times dom(B)$. A two-way cross-tabulation of the data determines all cell frequencies, n_{jk}'s. The Pearson chi-square test statistic, X^2, is defined as:

$$X^2 = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{jk} - \hat{m}_{jk}^{(0)})^2}{\hat{m}_{jk}^{(0)}},$$

where n_{jk} is the observed frequency count for the cell (j,k);

$\hat{m}_{jk}^{(0)}$, the expected count for the cell(j,k), assuming the validity of independence hypothesis for A and B. That is,

$\hat{m}_{jk}^{(0)} = \frac{n_{j+}\, n_{+k}}{n}$; where n_{j+} and n_{+k} are the jth row total and kth column total respectively of the two-way contingency table; and n denotes the total number of observations.

Based on validity of null hypothesis of independence, the statistic X^2 is asymptotically distributed as a chi-squared random variable with $(J-1)(K-1)$ degrees of freedom. Large observed values of X^2 or alternatively small p-values, do not support the null hypothesis of independence. A chosen $p-value = \alpha \leq 0.05$ leads to rejection of

independence hypothesis where α is the significance level of the test, see Christensen (1997).

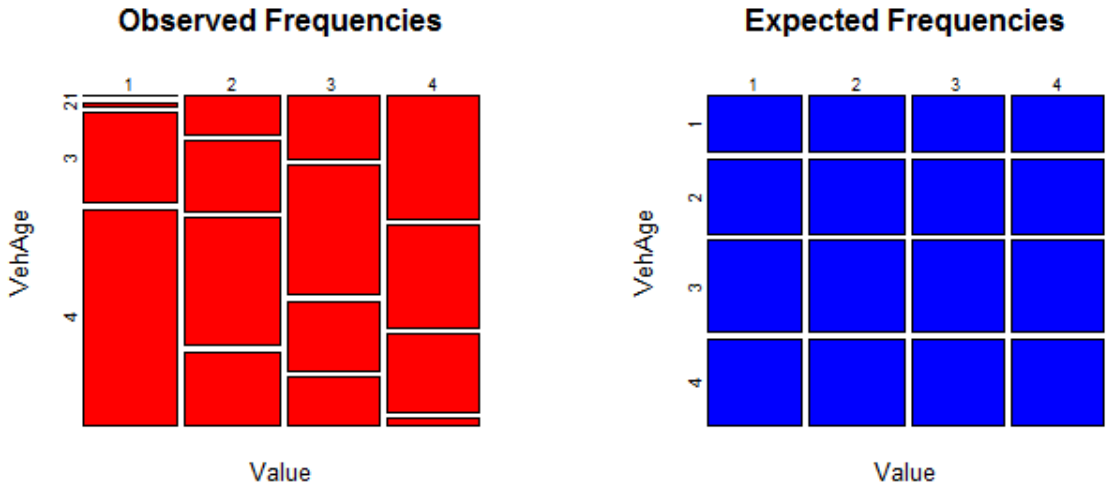
Our analysis of the Australian Automobile data involved seven categorical variables. Twenty one Pearson chi-square tests of independence were performed. It is interesting to note that only in the case of Claim and Gender, we failed to reject the null hypothesis of independence.

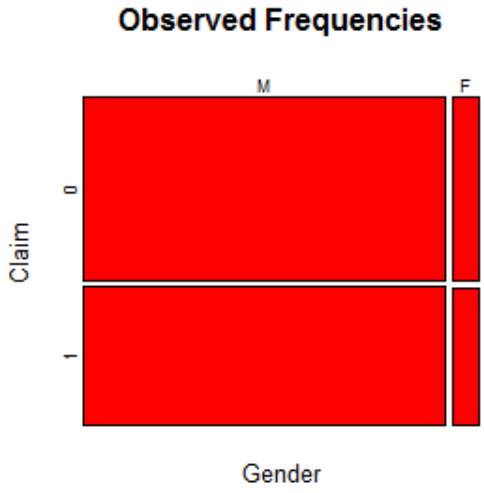
A Mosaic plot, see top-left panel of Exhibit 3.2, provides a graphical display corresponding to a two-way contingency table. The methodology to construct mosaic plots has been explained by Friendly (1994). A mosaic is composed of tiles, where each tile corresponds to a cell of contingency table. To the cell $(j,k),\ 1 \le j \le J, 1 \le k \le K$, there corresponds a tile labeled (j,k) whose width is proportional to n_{+k} and its height is proportional to $n_{j|k} = \frac{n_{jk}}{n_{+k}}$, the conditional count, see top-left panel of Exhibit 3.2, where $J = K = 4$. The area of the (j,k) tile is proportional to the cell frequency n_{jk}.

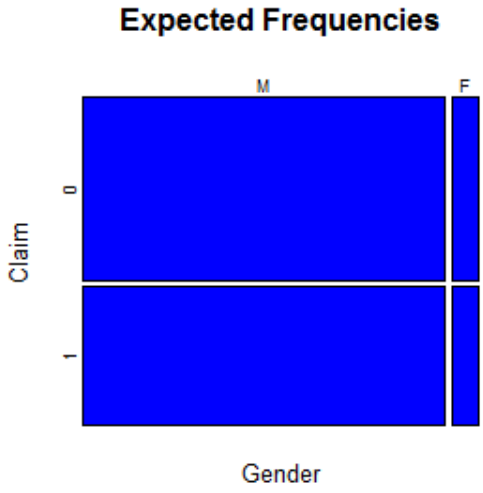
Under the independence assumption, the expected frequency for the cell (j,k) is $\hat{m}_{jk}^{(0)} = \frac{n_{j+}\, n_{+k}}{n}$. Hence, the height of the (j,k) tile is proportional to n_{j+} which does not depend upon k, the second variable. Based on the independence assumption for each j (horizontal level, representing the first variable) all tiles in the j^{th} row with differing k values have the same height, see the top-right panel of Exhibit 3.2.

The mosaics corresponding to observed and expected frequencies for categorical variables Value and Vehicle Age, Exhibit 3.2 top-left and top-right, are dissimilar. This lack of similarity is consistent with rejection of the chi-square test. On the other hand, the mosaics in the bottom-right and bottom-left of Exhibit 3.2 appear similar, and this outcome is consistent with failing to reject the independence test for Claim and Gender.

Exhibit 3.2: **Mosaic plots for Value & Vehicle Age as well as Claim & Gender**







Furthermore, one notices the anomaly between observed and expected frequencies for the cell corresponding to the Value of 1 and Vehicle Age of 4 in Exhibit 3.2. Hence, mosaic plots reveal further information beyond independence.

There are some limitations to the relying only on two categorical variables. First, if there are more than two categorical variables available, then it seems logical to use all the available variables, as

more information tends to lead to better decisions. Second, when there are other categorical variables available, say three in total, then in some instances, the inferences based on two categorical variables, a marginal approach, may contradict the conclusion derived based on using all three. These anomalies are referred to as Simpson's paradox, see Agresti (2002).

4. **Log-Linear models, Conditional Independence, and Graphical Modeling**

In this section we consider Log-Linear models, three-way mosaic plots, the concept of conditional independence, and graphical models as they relate to three categorical variables.

To perform a test of independence involving three categorical variables, one approach is to extend the Pearson chi-square test to the case of three factors. Alternatively, a preferred approach is to use the Log-Linear models.

Log-linear models are a special class of the Generalized Linear Models, GLM, which extend the classical regression models. In classical regression analysis, the mean of a continuous response variable is related to a set of explanatory variables assuming that the response variable is normally distributed. With Log-Linear models, one relates the expected cell count of a multidimensional contingency table to a set of categorical variables by specifying their main and interaction effects. This approach mirrors the ANOVA procedure where a continuous response variable is related to a set of explanatory factors. An advantage of using Log-Linear approach to the Pearson chi-square test is, it allows for testing alternative dependency structure among the categorical variables. Useful references for Log-Linear models are Agresti (2002), Fienberg (1980) and Christensen (1997).

We begin by describing Log-Linear models for two categorical variables A and B, although our main focus is with more than two variables. Let A and B have respective domains $dom(A) = \{1,2,\ldots,J\}$ and $dom(B) = \{1,2,\ldots,K\}$. For the cell $(j,k) \in dom(A) \times dom(B)$, related entities of interest p_{jk}, n_{jk}, and $m_{jk} = n\,p_{jk}$, denoting respectively the probability, the observed frequency count, and the expected count associated with the cell (j,k); n denotes the total number of observations in the data set. For a sample of size n, the random vector (A,B) has a Multinomial distribution. Multinomial distribution serves as the principal multivariate distribution for a random vector whose components are

categorical variables. Multinomial distribution is not as restrictive as Multivariate Normal which is used for random vector whose components are continuous random variables.

The Log-Linear model for two categorical variables A and B is specified as:

$$\log(m_{jk}) = u + u_j^A + u_k^B + u_{jk}^{AB}, \qquad (1)$$

where m_{jk} is the expected number of observation for the cell $(j,k), 1 \le j \le J \; and \; 1 \le k \le K$;

u, a constant term (an intercept),

u_j^A and u_k^B are the main effect terms due to A and B respectively,

and $u_{jk}^{AB}'s$ denote the two-factor interaction terms.

Testing for independence of A and B is equivalent to testing for the null hypothesis $H_0 : u_{jk}^{AB} = 0$ in model (1) above.

When the independence assumption prevails, model (1) reduces to

$$\log(m_{jk}) = u + u_j^A + u_k^B, \qquad (2)$$

Model (1) is referred to as the saturated model, and model (2) is referred to as the independence model. The saturated model is over-parameterized, i.e., the number of parameters u, $u_j^A, u_k^B, and \; u_{jk}^{AB}'s$ exceed the value of JK, the number of cells. To fit model (1) to cell frequencies, it is necessary to impose some restrictions on the number of parameters used. There are three alternative ways to impose restrictions on the parameters referred to as sum, treatment, and Helmert constraints, see references by Faraway (2005), Christenson(1997), or Feinberg (1980).
The R program uses the treatment constraint as default. The treatment constraint approach selects one level of A and one level of B as fixed and refers to them as base levels. The following identity shows that with the treatment constraints applied, the number of parameters used in the saturated model (1) is sufficient and not over specified:

1 + (J-1) + (K-1) + (J-1)(K-1) ≡ JK

Since the saturated model has as many parameters as there are cells in the cross-tabulated data, it provides a perfect fit to the cell frequencies and thus it provides no simplification in the context of modeling. The saturated model serves the purpose of being the base

model for comparing other simpler models to it, for example by comparing the independence model (2) to the saturated model (1).

To compare model (1) with model (2), the appropriate test statistic is the Deviance statistic, G^2

$$G^2 = 2\sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk} \log\left(\frac{n_{jk}}{\hat{m}_{jk}^{(0)}}\right)$$

where $\hat{m}_{jk}^{(0)}$ is the expected count for the cell (j,k) assuming the independence model (2) is valid, i.e, $\hat{m}_{jk}^{(0)} = \frac{n_{j+}\, n_{+k}}{n}$ is the Maximum Likelihood Estimator for m_{jk}, expected cell count, under the independence assumption.

The test statistic G^2 is asymptotically distributed as a chi-squared random variable with $(J-1)(K-1)$ degrees of freedom, see Christenson(1997) for details.

The R output for testing the independence of Claim and Location is given in Exhibit 4.1 below:

Exhibit 4.1 R Output for Testing the Independence of Claim and Location

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.629079   0.007898 1219.14   <2e-16 ***
Claim1      -2.615550   0.015234 -171.69   <2e-16 ***
LocationB   -0.201059   0.011673  -17.22   <2e-16 ***
LocationC    0.230473   0.010488   21.98   <2e-16 ***
LocationD   -0.691065   0.013552  -50.99   <2e-16 ***
LocationE   -1.014917   0.015181  -66.86   <2e-16 ***
LocationF   -1.517097   0.018461  -82.18   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 79977.618  on 11  degrees of freedom
Residual deviance:    18.117  on  5  degrees of freedom
AIC: 147.66
```

The Deviance (Residual deviance), has a value of 18.117. This statistic is used to test the independence of Claim and Location. Its asymptotic distribution is a chi-squared distribution with 5 degrees of freedom, the difference between the number of parameters used in model (1) and (2). It has a p-value of 0.00280 which implies that we reject the independence hypothesis for Claim and Location in this instance. This result is consistent with the Pearson Chi-squared test performed in section 3. Recall that the only case where Pearson Chi-squared tests of independence failed was for Claim and Gender.

We now introduce the concept of graphical models for the case of two categorical variables. Further elaboration on this subject is given below when we have more than two variables.

Graphical models are used to illustrate relationships among a number of variables. In the case of two categorical variables A and B, the situation is relatively simple: either A is independent of B or they are not independent. Exhibit 4.2 below illustrates this viewpoint.

Exhibit 4.2: Graphical Presentation of Claim & Location and Claim & Gender

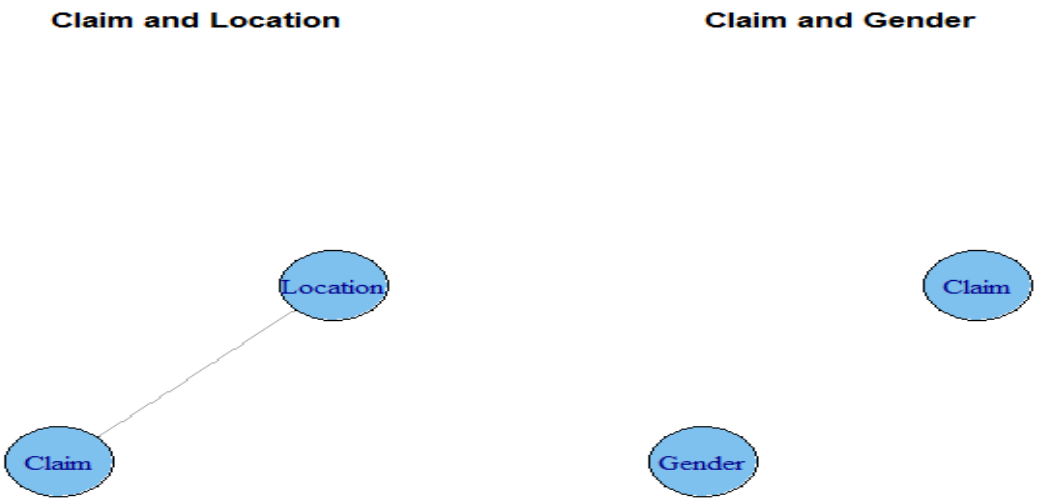


Exhibit 4.2 presents two graphs next to each other. The vertices (nodes) represent the variables. The presence of an edge (chord)

between two nodes implies the variables are related. Absence of an edge implies independence.

To summarize, we can test for independence using either the Pearson chi-square test X^2, section 3, or use Deviance G^2 as defined above for the Log-Linear model. Furthermore, we can represent our results by graphs as shown in Exhibit 4.2.

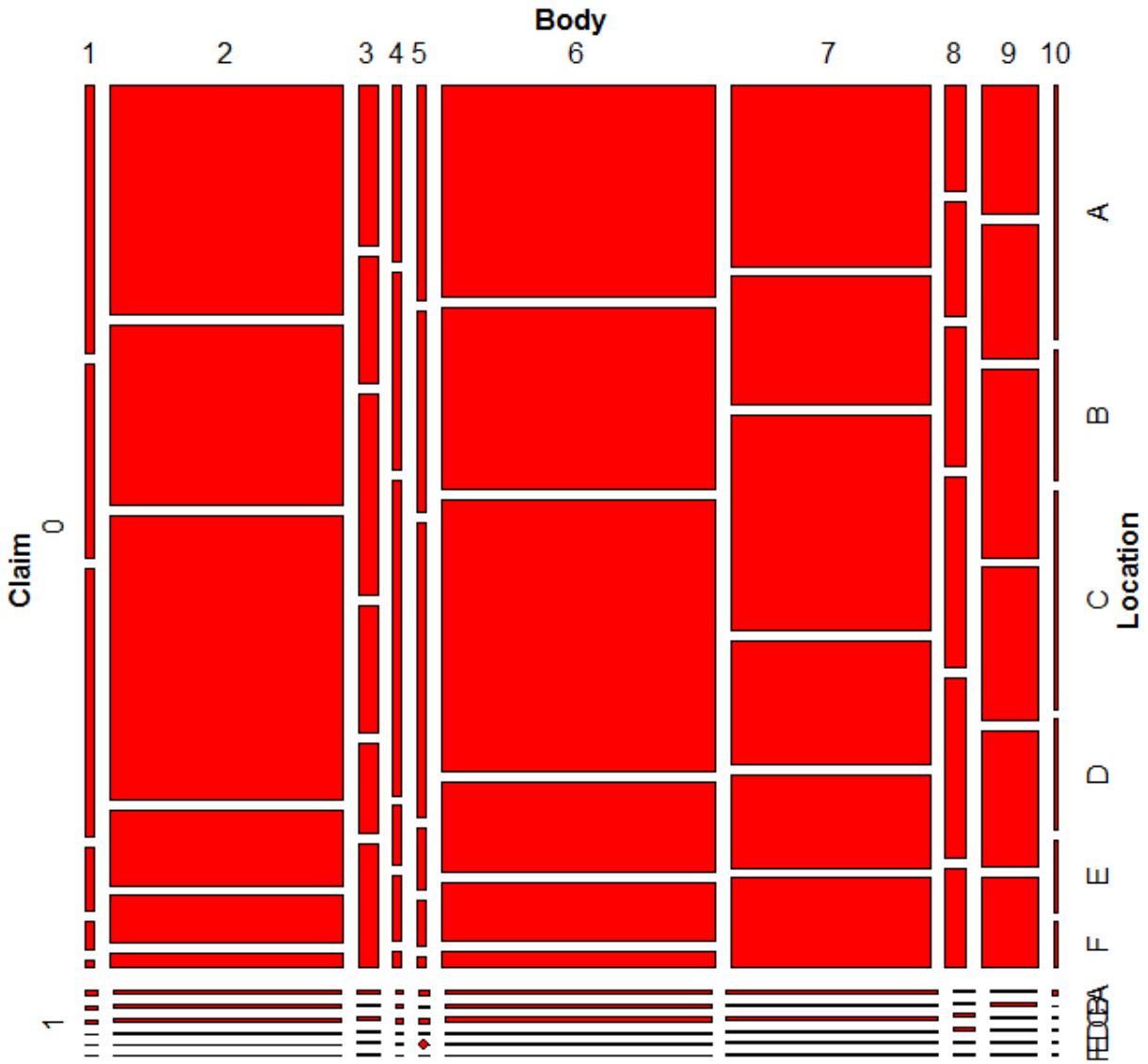
Next, we proceed to describe mosaic plots for visualizing a three-way contingency table. Table 4.3 summarizes the data for categorical variables Claim, Body and Location in the form of three-way contingency table derived from the Australian automobile insurance data.

Table 4.3: Three-way Contingency Table for Claim, Body and Location

	Body	1	2	3	4	5	6	7	8	9	10
Claim	Location										
0	A	230	4874	280	143	179	5284	3281	208	670	78
	B	166	3810	221	160	167	4559	2330	225	698	40
	C	230	6035	352	256	244	6780	3907	275	982	67
	D	55	1612	222	48	52	2239	2246	374	795	34
	E	24	1009	157	53	39	1465	1696	354	707	22
	F	7	311	217	14	9	430	1628	194	474	14
1	A	23	332	27	11	21	367	250	7	38	9
	B	23	299	28	12	13	323	197	15	52	3
	C	16	434	29	14	20	520	294	33	48	4
	D	4	110	18	2	6	130	149	27	45	5
	E	1	65	13	2	0	108	123	23	48	3
	F	1	24	15	2	2	28	160	15	29	4

Exhibit 4.4 displays a mosaic plot corresponding to Table 4.3.

Exhibit 4.4: Mosaic Plot for Claim, Body and Location



The construction of mosaic plot for multidimensional contingency table, three categorical variables in this instance, is based on the exposition given by Friendly (1994).

For three categorical variables A, B, and C, with typical levels of j, k, and l respectively, to each cell (j,k,l) of the three-way contingency table there corresponds a tile (j,k,l) constructed by the following the three sequential steps:

Step 1) For A, the first categorical variable, create vertical strips with width proportional to n_{j++}.

Step 2) Each vertical strip, in Step 1) is subdivided horizontally with height proportional to $\frac{n_{jk+}}{n_{j++}}$, conditional count of the second variable B given the first variable A.

Step 3) Each JK rectangle in Step 2) is further subdivided vertically with widths proportional to $\frac{n_{jkl}}{n_{jk+}}$.

In this fashion a tile constructed in Step 3), labeled as (j,k,l), has an area proportional to n_{jkl}.

The three steps involved above is analogous to writing the joint probability of the cell (j,k,l) as product of marginal and conditional probabilities as expressed by

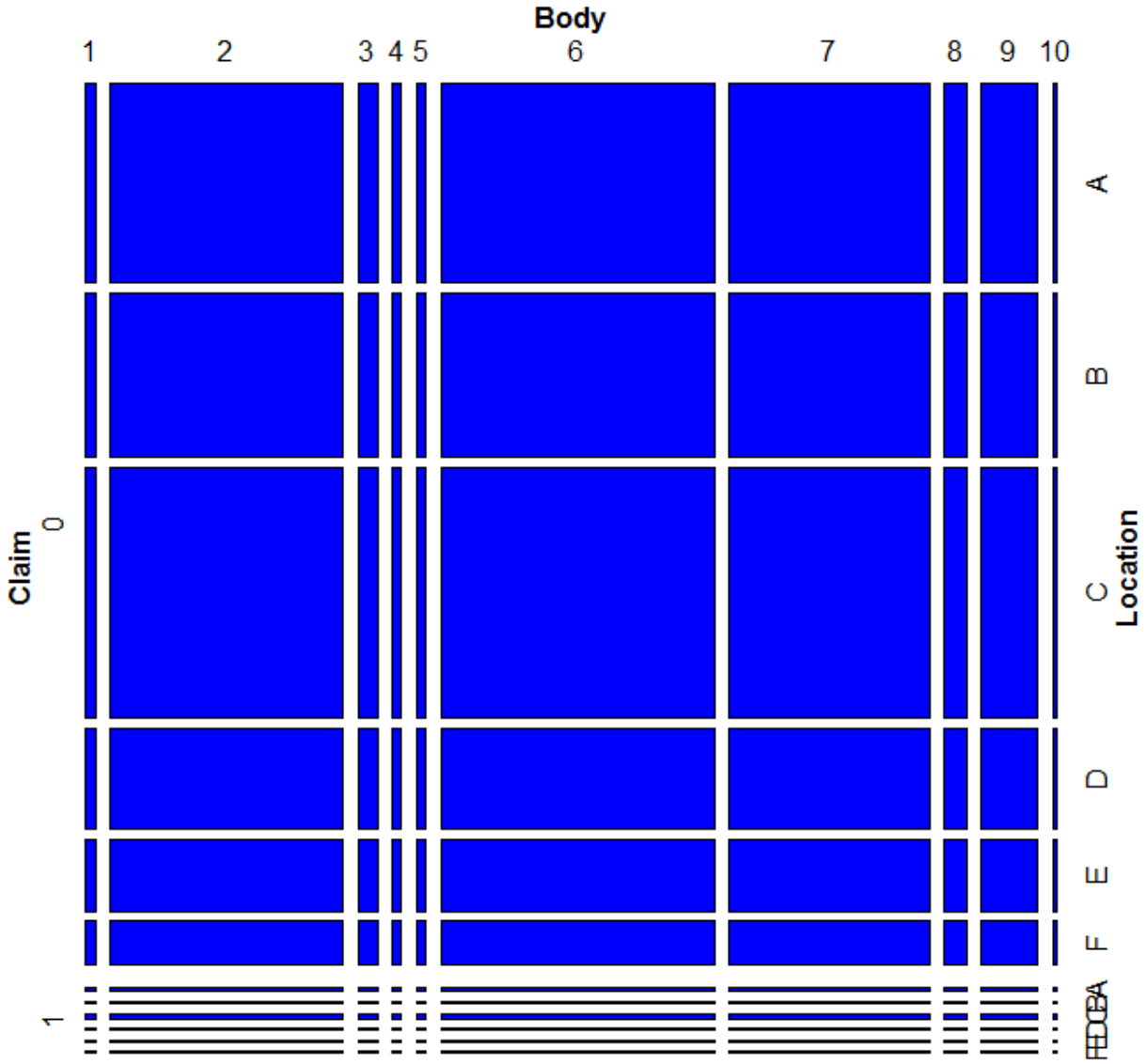
$$P(A=j,B=k,C=l)=P(A=j)P(B=k\mid A=j)P(C=l\mid A=j,B=k)$$

Exhibit 4.5 shows the mosaic plot for the three categorical variables Claim, Body and Location based on the expected frequency under the assumption of independence. That is the joint probability of the cell (j,k,l), p_{jkl} is computed as product of three marginal probabilities, i.e., $p_{jkl}=p_{j++}\,p_{+k+}\,p_{++l}$. All probability items are estimated by appropriate cell count ratios.

Applying Log-linear models, the independence assumption was not supported for any of three categorical variables studied. Hence, it should not be surprising to see Exhibits 4.4 and 4.5 having different appearances.

As the number of categorical variables increases, then it becomes harder to interpret the pattern of dependency among variables based on mosaic displays. This is due to addition of a new variables requires further subdivision of each existing tile.

Exhibit 4.5: Mosaic Plot for Claim, Body and Location Based on the Hypothesis of Independence



Now, we consider Log-Linear models for the case of three categorical variables A, B, and C. The notations previously used will be extended to the case of three.

A Log-Linear model for three categorical variables A, B, and C, with C having a typical value of l, where $l \in dom(C) = \{1,2,\ldots,L\}$, is defined by

equation (3) below. The model in (3) is referred to as the saturated Log-Linear models for three categorical variables.

$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jk}^{AB} + u_{jl}^{AC} + u_{kl}^{BC} + u_{jkl}^{ABC}, \qquad (3)$$

where m_{jkl} is the expected count for the cell (j,k,l), with

$(j,k,l) \in dom(A) \times dom(B) \times \dim(C)$;

u is the constant (intercept) term,

$u_j^A, u_k^B, and\ u_l^C$ are the main effect terms, and

$u_{jk}^{AB,}, u_{jl}^{AC}, u_{kl}^{BC}, and\ u_{jkl}^{ABC}$ are two-factors and three-factors interaction terms.

For a sample of size n, the random vector (A,B,C) has a multinomial distribution. The number of parameters in model (3) is over-specified and subject to constraint as in the case of two categorical variables above.

The saturated model (3) is the largest model with respect to three categorical variables. By excluding certain u-terms in (3) above, alternative dependency structures among A, B, and C may be considered. The saturated model serves as a base model for comparing to other parsimonious Log-Linear models.

There are two subclasses of Log-Linear models which are of interest to us: the **hierarchical** models and the **graphical** models. The definition of these terms is based on the one given by Edwards (2000). In a hierarchical Log-Linear model, if a u-term is excluded (set equal to zero) then all its higher-order related u-terms are also excluded. With regard to graphical models, a subclass of hierarchical models, are formed by excluding a set of two-factor interaction terms (and hence their higher-order related terms). For a more detailed discussion of these concepts refer to Lauritizen (1996).

The independent model, the smallest Log-Linear model with all the three categorical variables included, is presented as:

$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C \qquad (4)$$

Among Log-linear models, the **conditionally independent** models are of much interest. These models are not as detailed as the saturated model, but provide additional dependency structure not provided by the independence model. They are easier to interpret and belong to the class of graphical models.

The conditional independence property, as it relates to three categorical variables A, B, and C, is defined as follows:
A and B are conditionally independent given C if

$$P(A = j, B = k \mid C = l) = P(A = j \mid C = l)\, P(B = k \mid C = l),$$

for all $(j,k,l) \in dom(A) \times dom(B) \times \dim(C)$.
The notation used to express conditional independence, is $A \underline{\perp\!\!\!\perp} B \mid C$ due to Dawid(1979).

The three conditional independence models of interest are:

$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jl}^{AC} + u_{kl}^{BC} \qquad (5a)$$

$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jk}^{AB} + u_{kl}^{BC} \qquad (5b)$$

$$\log(m_{jkl}) = u + u_j^A + u_k^B + u_l^C + u_{jk}^{AB} + u_{jl}^{AC} \qquad (5c)$$

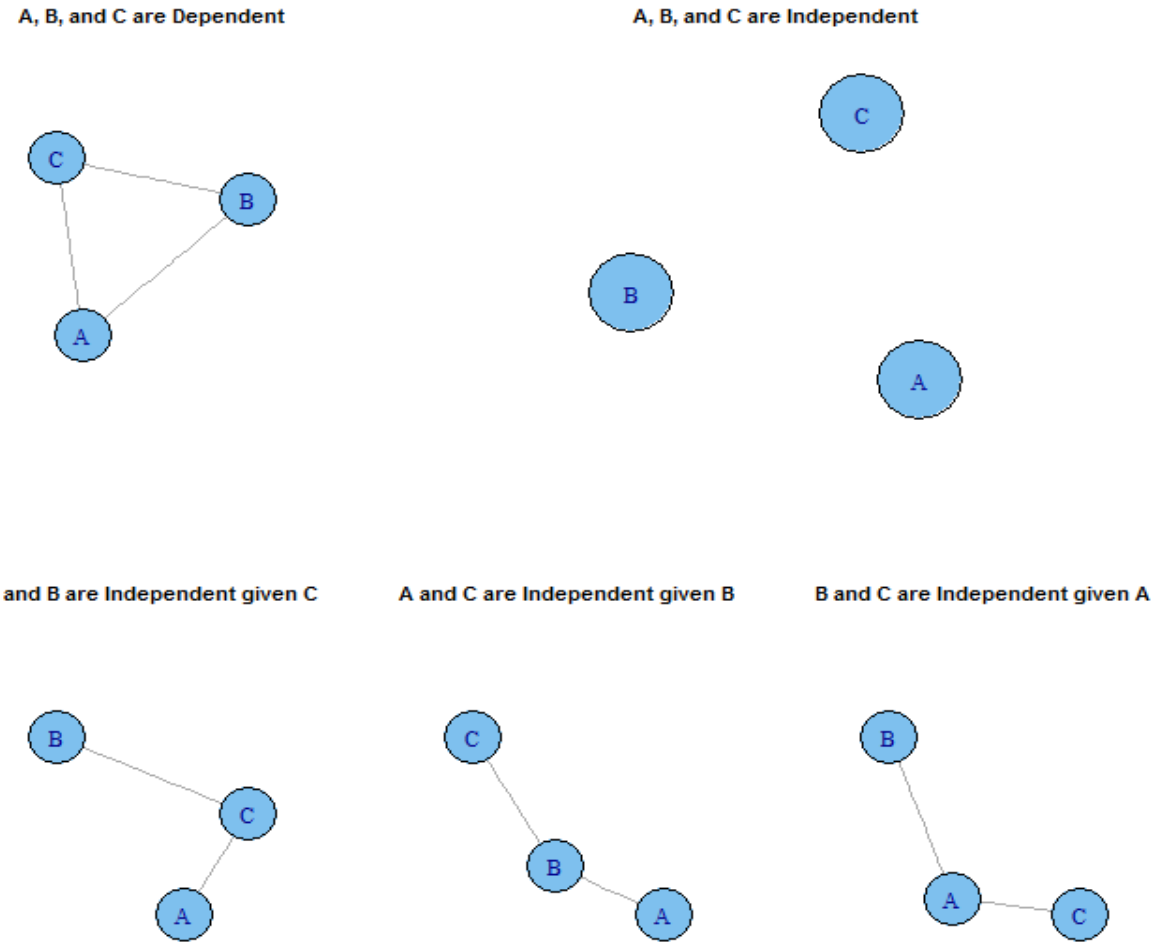
Using the conditional independence notation, we have $A \underline{\perp\!\!\!\perp} B \mid C$, $A \underline{\perp\!\!\!\perp} C \mid B$, and $B \underline{\perp\!\!\!\perp} C \mid A$ corresponding to (5a), (5b), and (5c) respectively.

In order to discuss further the graphical models, we need to introduce some basic elements of the **graph theory**. Graph theory is a branch of mathematic, see Berge (2001). Graph theory has been applied to transportation networks, and social networks, Kolaczyk (2009); and used in computer science, see Cook and Holder (2007). For interaction of graph theory and Log-Linear models refer to the following references: Whittaker (1990), Lauritzen (1996), and Edwards (2000).

A graph G, is a pair (V,E) where V is a finite set of *vertices* and E, a subset of $V \times V$, is a set of *edges*. Here, the vertices represent variables, and for two vertices a and b such that $(a,b) \in E$ implies a relationship between vertices a and b. If $(a,b) \in E$ then we say that vertices a and b are *adjacent* to each other. A graph is *simple if we* exclude loops and multiple edges. A graph is undirected if $(a,b) \in E$ implies $(b,a) \in E$. The graphs considered here are simple and undirected. A graph is *complete* if all of the vertices in the graph are adjacent to each other.

Exhibit 4.6 presents five graphs of interest corresponding to the models (3), (4) and (5) above.

Exhibit 4.6: Graphical Models corresponding to Complete, Independent, and Conditional Independence



The top left graph of Exhibit 4.6, corresponds to a graph that is complete, the saturated model of (3). The top right graph with no edges corresponds to the independent model of (4). The three graphs in the lower part of the exhibit, from left to right, correspond to conditional independence models of $A \perp\!\!\!\perp B \mid C$, $A \perp\!\!\!\perp C \mid B$, and $B \perp\!\!\!\perp C \mid A$ respectively, i.e., to (5a), (5b), and (5c).

Relative to the saturated model, a conditional independent model is a more parsimonious model in the sense that it requires fewer parameters to be specified. Furthermore, the conditional independent models are easier to explain and interpret graphically. For instance, the

conditional independent model of (5b), $A \perp\!\!\!\perp C \mid B$, implies that the edge AC has been removed from the complete model.

Returning to the Australian Auto data, with seven categorical variables, there are potentially 35 possible Log-Linear models involving 3 categorical variables.

Each of these 35 models failed the test of independence. Next, we considered the conditionally independent models, models (5a), (5b) and (5c), for 35 cases. Exhibit 4.7 provides the summary of testing for the conditional independence where the results were **statistically significant.**

	Exhibit 4.7: Results of Conditional Independent Tests
1	Claim and Gender are conditionally independent of Value
2	Claim and Gender are conditionally independent of Body
3	Claim and Body are conditionally independent of Driver Age
4	Claim and Gender are conditionally independent of Vehicle Age
5	Claim and Gender are conditionally independent of Location
6	Claim and Gender are conditionally independent of Driver Age
7	Claim and Location are conditionally independent of Driver Age

The function ciTest_table()of the R package gRim, see Hojsgaard, Edwards, and Lauritzen (2012), was used to perform the conditional independent tests. With three categorical variables, the implication of conditional independence test is a tantamount to removal of an edge from the respective complete graph. Exhibit 4.8 below provides the R output in the case of three variables: Claim, Gender, and Vehicle Age.

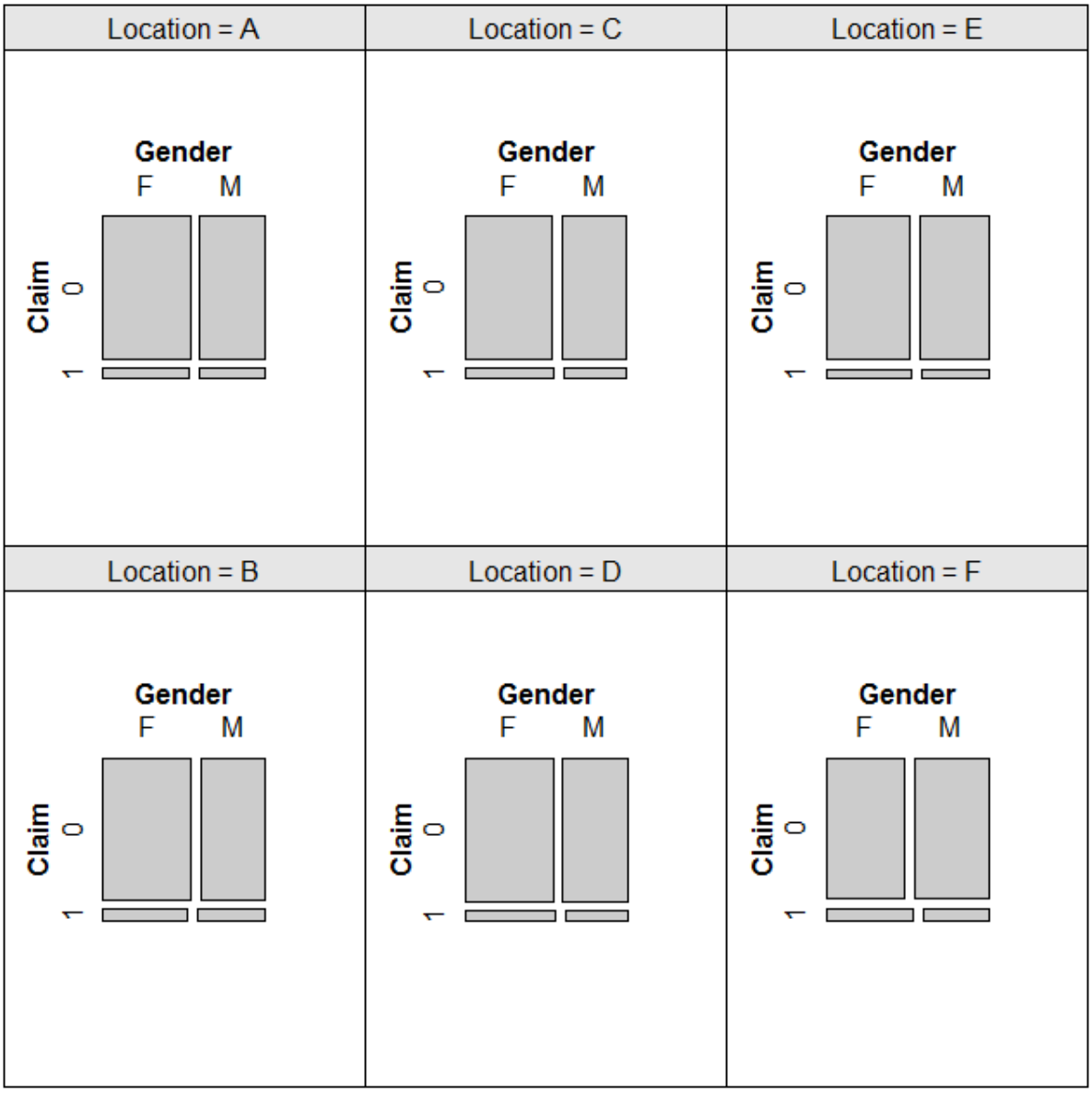
Exhibit 4.8: R Output for Testing Conditional Independence of Claim, Vehicle Age and Gender

```
# d.3.245.table denotes the three dimensional frequency table for
# categorical variables Claim, Vehicle Age and Gender.


# Results a) for Claim & Vehicle Age given Gender
ciTest_table(d.3.245.table, set=c("Claim", "VehAge", "Gender"))
Testing Claim _|_ VehAge | Gender
Statistic (DEV):   27.246 df: 6 p-value: 0.0001 method: CHISQ


# Results b) for Claim & Gender given Vehicle Age
ciTest_table(d.3.245.table, set=c("Claim", "Gender", "VehAge"))
Testing Claim _|_ Gender | VehAge
Statistic (DEV):    1.225 df: 4 p-value: 0.8740 method: CHISQ


# Results c) for Vehicle Age & Gender given Claim
ciTest_table(d.3.245.table, set=c("VehAge", "Gender", "Claim"))
Testing VehAge _|_ Gender | Claim
Statistic (DEV):  288.004 df: 6 p-value: 0.0000 method: CHISQ
```

In Exhibit 4.8, the p-value for testing the conditional independence of Claim and Gender given the Vehicle Age is 0.8740. It has the implication that test fails to reject the conditional independence test in this instance. Other results in Exhibit 4.7 were similarly derived.

Before proceeding to the next section involving more than three categorical variables, it is worth introducing an additional exploratory tool for visualizing “conditional” relationship among three variables. Exhibit 4.9 displays the relationship between Claim and Gender for each level of Location variable.

Exhibit 4.9: Conditional Plot of Claim and Gender Given Location



Reviewing the six panels of Exhibit 4.9, once notices the similarity of these panels with respect to Claim and Gender for each level of Location. This observation is consistent with formal test of conditional independence, case 5 of Exhibit 4.7.

5. Graphical Models.

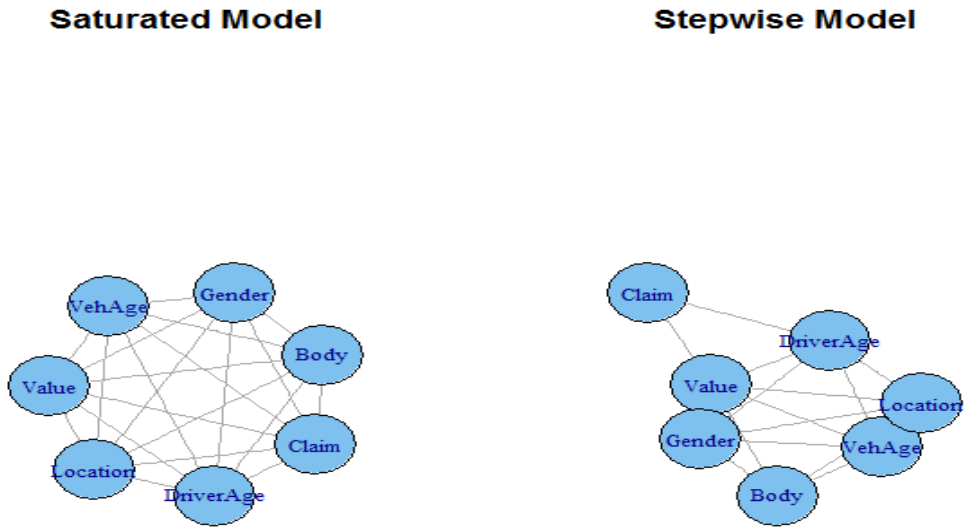
Log-Linear models of the previous section can be extended to more than three categorical variables. By increasing the number of variables, one encounters two kinds of problems. The first problem is referred to as the “model selection”, and the second as the “estimation” problem. Let us elaborate on each of these two issues. By excluding certain two-factor, three-factor or other higher order factor interaction terms, different Log-Linear models are obtained. The number of potential Log-Linear models increases in a non-linear fashion as the number of categorical variables increases. According to Christensen (1997), with four categorical variables there are 113 ANOVA type models with their main effects included. Furthermore, with five categorical variables, there are several thousand models to choose from. The model search space, the set of potentially acceptable Log-Linear models, is very large, and finding an “optimum” model among the potential models is a challenging task. It may be more expedient to search among the smaller class of graphical models for a suitable candidate. Graphical models have the advantages of being presented as graphs and are subject to easier interpretation.

The second problem is related to estimation of parameters of Log-Linear models. Some data sets have cells with zero frequency. In these situations, some software may either fail to fit a model to the data, or it may make adjustments and proceed to fit a model accordingly. The handling of empty cells is not uniform among different software. Hence, there are uncertainties with regard to the output of the software selected, see Hojsgarrd, Edwards and Lauritzen (2012). Obtaining empty cells is not that uncommon when the number of categorical variables increases. With a fixed sample size, the observations need to be spread to a larger number of cells resulting in some cells being empty.

The approach in this paper for selecting a model is as follows. First, we fit a saturated model to the data using all the available categorical variables. The fitted saturated model corresponds to a complete graph. Here, we use the term model and graph interchangeably. The second step involves using the fitted saturated model as an input to the stepwise function of R in the gRim package. The output of the stepwise function is a more parsimonious fitted Log-Linear model belonging to the class of graphical models. The stepwise function uses a backward elimination procedure by removing certain edges of the saturated graph, thus producing a pruned subgraph; see Hojsgarrd, Edwards and Lauritzen (2012).

The graphs of the saturated fitted model as well as the graph produced by applying the stepwise function, here referred to as the stepwise model, are shown in Exhibit 5.1.

Exhibit 5.1: Graphical Models for Saturated and Stepwise Models Based on the Australian Auto Data



An insight into structure of the stepwise graph is provided by examining its *cliques*. A clique is a complete (maximal) *subgraph* such that by enlarging its vertex set, it would lose the property of being complete.

An intuitive characteristic of a clique based on the exposition of Hanneman and Riddle (2005), is as follows: ***a clique is a subgraph of a graph whose nodes are more closely and intensely related to one another than they are to other nodes of the graph.***

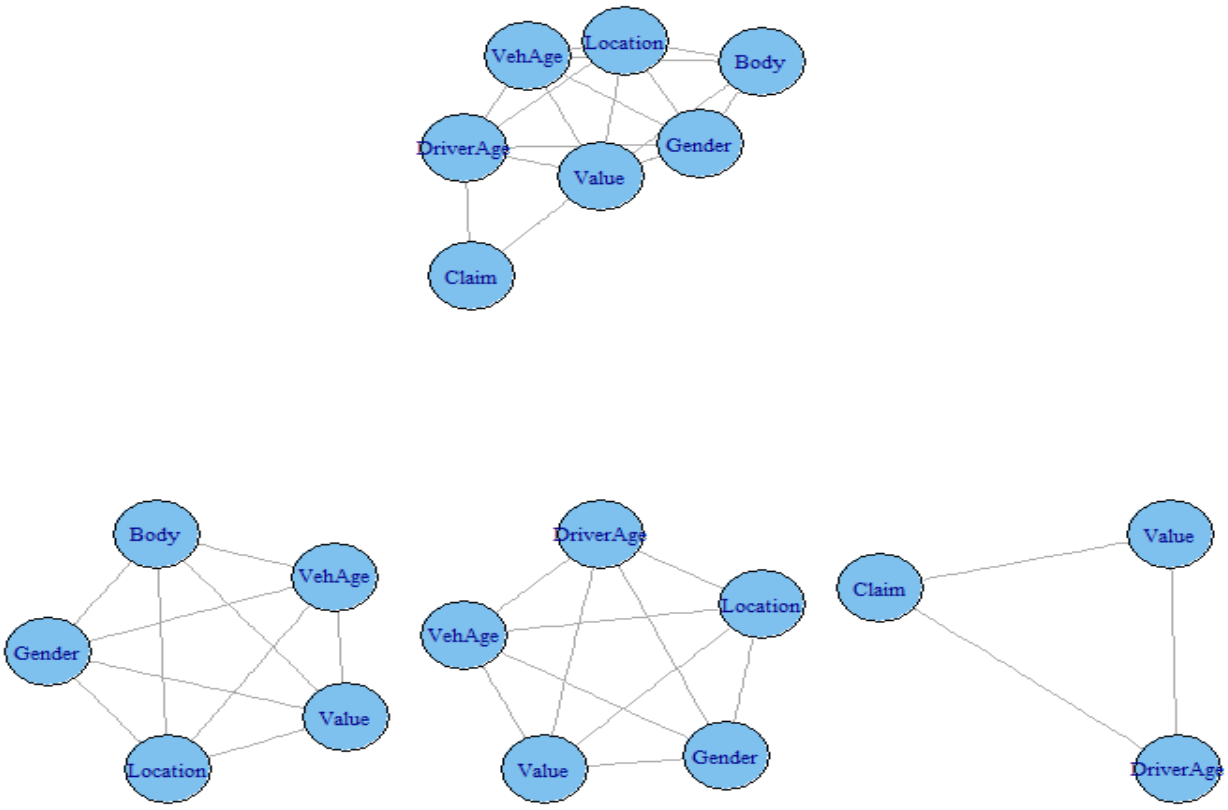
Exhibit 5.2 illustrates the composition of three cliques associated with the graph of the fitted stepwise model.

Exhibit 5.2: Cliques of the Fitted Stepwise Model

```
RBGL::maxClique(as(model.step,"graphNEL"))
$maxCliques
$maxCliques[[1]]
[1] "Value"    "VehAge"   "Gender"   "Location" "Body"
$maxCliques[[2]]
[1] "Value"     "VehAge"    "Gender"    "Location"  "DriverAge"
$maxCliques[[3]]
[1] "Value"   "Claim"   "DriverAge"
```

Exhibit 5.3 illustrates the fitted stepwise model and its cliques.

Exhibit 5.3: Stepwise Graph and Its Cliques



Two of the cliques of Exhibit 5.3, provide exposure (underwriting) information and third clique, the lower right, provides claim information. Two of the cliques show a strong **binding** among the four variables Value, Vehicle Age, Gender and Location. By confining ourselves to graphical models, and examining their cliques, we can gain an insight into variables that are closely related.

Now, we give an example for reviewing claim frequency rates which utilizes the information about the cliques in Exhibit 5.2. Let us compute average Claim occurrence rate for each cell in three circumstances. The three cases depend upon which underwriting factors have been selected. We shall refer to these cases as “All”, “Clique 1”, and “Clique 2”. Table 5.1 gives summary of the statistics produced. We shall explain the statistics corresponding to the “All” case. The figures for the other two cases were similarly derived.

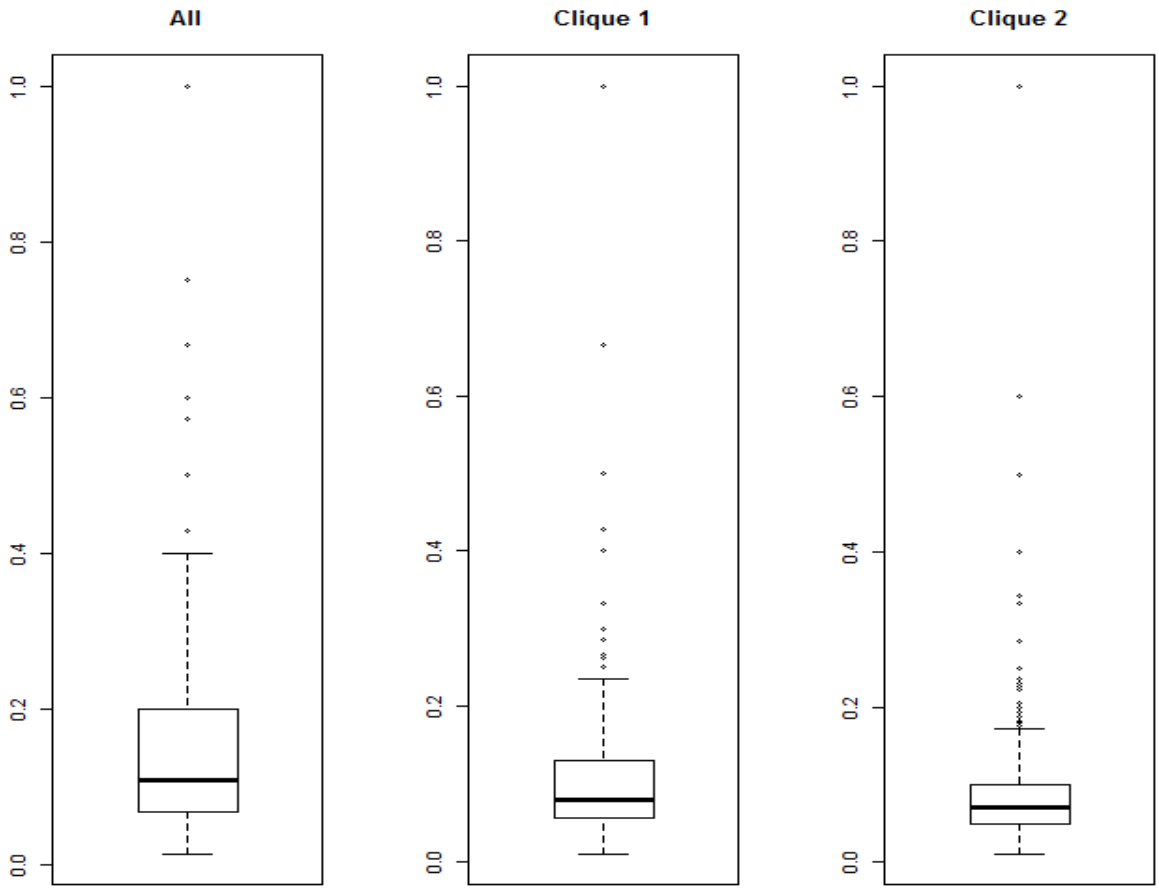
For the Australian auto insurance, the “All” case involved the six underwriting variable. The variables involved were Value (after conversion from numeric to ordinal), Body (after reducing the number of level from 13 to 10), Vehicle Age, Gender, Location, and Driver Age. The number of potentially distinct cells arising from different combination of levels of these six variables was **11,520**(4*10*4*2*6*6). The actual number of non-empty observed cells was 5,063. The average value of Claim occurrence for each cell, referred to as rate, was computed. Only 1,905 cells had **positive** average Claim occurrence rates. These positive claim rates may be utilized for computing or reviewing frequency rates. Note that the traditional ratemaking procedures may not consider some of these 1,905 cell rates as credible due to low cell counts. With regard to these positive average Claim occurrence rates, we computed the values of Minimum, Q1 (first quantile), Median, Mean, Q3 (third quartile) and Maximum, see Table 5.1. The figures for “Clique 1” and “Clique 2” cases were similarly derived.

Table 5.1- Review of Claim Frequency based on Cliques

ID	Variables	Potential # cells	Observed # cells	# cells with positive "freq."	Min.	Q1	Median	Mean	Q3	Max.
All	Value (4) Body (10) Vehicle Age (4) Gender (2) Location (6) Driver Age (6)	11,520	5,063	1,905	0.0135	0.0667	0.1071	0.1869	0.2000	1.0000
Clique 1	Value (4) Body (10) Vehicle Age (4) Gender (2) Location (6)	1,920	1,211	701	0.0099	0.0562	0.0800	0.1275	0.1290	1.0000
Clique 2	Value (4) Vehicle Age (4) Gender (2) Location (6) Driver Age (6)	1,152	1,047	834	0.0111	0.0500	0.0702	0.0875	0.0997	1.0000

For the above three cases, we constructed boxplots in order to compare the distribution of positive average Claim occurrence rates, see Exhibit 5.4.

Exhibit 5.4 Comparison of Average Claims Rates



Reviewing the boxplots, we note that “frequency” rates for the “Clique 2” are more stable. There are less volatile, and the median and mean values are close to each other. For “All” and “Clique 1” cases, the average Claim occurrence rates may well be influenced by outliers. It is difficult to visually see the number of outliers appearing in Exhibit 5.4. I used the “out” attribute of the R boxplot function for determining outliers (extremes), see the following link: http://stat.ethz.ch/R-manual/R-devel/library/grDevices/html/boxplot.stats.html

The outlier definition used is associated with the “out” attribute of the boxplot function defined as “the values of any data points which lie beyond the extremes of the whiskers”. The percentage of outlier values for the “All”, “Clique 1” and “Clique 2” were 10.4, 11.1, and

5.2 respectively. Our example may help in determining or reviewing a collection of “frequency” rates which are more **stable** and are based on **smaller** number of rating factors.

Finally, we shall discuss briefly the notion of overfitting as it applies here. The graphs in Exhibit 5.3 were based on the using the entire data set. Results based on using all data may be too “optimistic”--close to the data--and may not necessarily generalize well to similar unseen data. This phenomenon is referred to as overfitting. Tan, Steinbach and Kumar (2006) discuss overfitting as it applies to classification, a supervised learning task. As was mentioned above, our study is mainly an exercise in unsupervised learning, and moreover, we did not have access to similar additional unseen data. So, to validate our findings with regard to the stepwise graph and its cliques (Exhibit 5.3), we considered replicating our results by selecting 10 random samples of equal size from the original data set. For each random sample: (1) we determined the fitted saturated model, (2) used the sample fitted saturated model as input to the stepwise function to determine the corresponding fitted stepwise model, and (3) examined the cliques associated with fitted stepwise model for each of the 10 samples. The result of these replications is summarized in Table 5.4 below.

Table 5.2: Replication of Graphical Modelling Based on Ten Random Samples

Cliques based on the entire data set

Body	Gender	Location	Value	VehAge
DriverAge	Gender	Location	Value	VehAge
Claim	DriverAge	Value		

Cliques with 4 variables

				Freq.	Sample ID
Body	Gender	Value	VehAge	4	2, 4, 7, 9
DriverAge	Gender	Value	VehAge	2	2, 9

Cliques with 3 variables

Location	Value	VehAge	10	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Body	Value	VehAge	6	1, 3, 5, 6, 8, 10
DriverAge	Value	VehAge	5	3, 5, 6, 7, 10
Body	Gender	Value	3	1, 8, 10
DriverAge	Gender	Value	3	1, 4, 8
Body	Gender	VehAge	1	5
Claim	DriverAge	Gender	1	1
Claim	Gender	Value	1	10

Cliques with 2 variables

Claim	Value	4	2, 3, 6, 7
Body	Gender	2	3, 6
Claim	DriverAge	2	5, 9
Claim	Gender	1	8
Claim	VehAge	1	4

Reviewing the results in Table 5.4, we note that using the entire data set produced two cliques of size 5 which do not appear among the cliques produced by the 10 replications. This result may be attributable to the *size* effect. It is conceivable that a pattern observed in a large data set may not reveal itself in smaller data sets, see BIG DATA by Mayer-Schonberger and Cukier (2013). But the two cliques of size 4 from the replicated samples share the same variables as the cliques of size 5 based on the entire data. It is interesting to note that the cliques involving the three variables Location, Value and Vehicle Age appeared in all 10 random samples. On the whole, the

results obtained by performing the replications do not appear to contradict the findings based on the entire data.

6. Summary and concluding remarks

In this paper, we studied a number of categorical variables related to an Australian automobile insurance data. We began by using Exploratory Data Analysis tools to understand the nature of our data. Transformations of some of the variables was done as a part of preprocessing the data. Visualization tools: bar charts, mosaics, and conditional plots were informally used. We constructed a number of multi-dimensional contingency tables for summarizing the information regarding the categorical variables used.

Tests of independence based on chi-square statistic and Log-Linear models were discussed. Concepts of conditional independence and graphical modeling were introduced. By examining the cliques of a parsimonious graphical model fitted to the data, one obtains insight into which combination of categorical variables tend to bind together. Furthermore, we discussed briefly issues related to model selection and overfitting. Finally, an example was given that utilized the information provided by cliques of a fitted graphical model with regard to dependency structure among rating variables. Such an analysis may supplement the traditional ratemaking reviews of classification rates.

Appendix A.1: A brief description of the Australian Automobile Data set

This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005. There are 67856 policies, of which 4624 (6.8%) had at least one claim.

Variables:

veh_value vehicle value, in $10,000s
exposure 0-1
clm occurrence of claim (0 = no, 1 = yes)
numclaims number of claims
claimcst0 claim amount (0 if no claim)
veh_body vehicle body, coded as
- BUS
- CONVT = convertible
- COUPE
- HBACK = hatchback
- HDTOP = hardtop
- MCARA = motorized caravan
- MIBUS = minibus
- PANVN = panel van
- RDSTR = roadster
- SEDAN
- STNWG = station wagon
- TRUCK
- UTE - utility

veh_age age of vehicle: 1 (youngest), 2, 3, 4
gender gender of driver: M, F
area driver's area of residence: A, B, C, D, E, F
agecat driver's age category: 1 (youngest), 2, 3, 4, 5, 6

Appendix A.2: R functions

Section	Subject	R function
2	Exhibit 2.1	str()
2	Exhibit 2.2	summary()
2	Quartiles	quantile()
2	Exhibit 2.3	dist()
2	Dendrogram	hclust(), plot()
2	Bar Chart	barplot
3	Chi-square test	chisq.test()
3	Exhibit 3.1	table()
3	Exhibit 3.2	mosaic()
4	Log-Linear fit	glm()
4	Exhibit 4.3	dmod(), plot(), stepwise()
4	Exhibit 4.5	ciTest_table()
5	Exhibit 5.2	RBGL::maxClique()

Appendix A.3: Preprocessing

The preprocessing of the data involved the following steps:

a) Variable names in the original data set were re-named for the sake of easier reference. The original names appear in the Appendix A.1 and re-produced here in parenthesis next adjacent to their names used here: Value (veh_value), Claim (clm), Body (veh_body), VehAge (veh_age), Gender (gender), Location (area), and DriverAge (agecat).

b) The variable vehicle value, a numeric variable, was transformed into an ordinal variable as Value.

c) Apart from change of names, there were no other changes concerning the five variables: Claim, VehAge, Gender, Location, and DriverAge.

d) The variable Body had originally thirteen levels presenting the Type of Body of the Vehicle. Due to low frequencies of some levels, it was decided to combine lower frequency levels, thus reducing the number of levels for analysis to 10.

e) Three other variables were excluded from this study.

References

Agresti, A. *Categorical Data Analysis*, Wiley (2002).

Berge, C. *The Theory Of Graphs*, Dover Publication (2001).

Cook D.J., and Holder L.B., *Mining Graph Data*, Wiley (2007).

Christensen R.C., *Log-Linear Models and Logistic Regression*, Springer (1997).

Dawid, A.P., "Conditional independence in statistical theory (with discussion)", *J. R. Stat. Soc. B* 41: 1-31, 1979

de Jong, P. and Heller, G. Z., *Generalized Linear Models for Insurance Data*, Cambridge University Press (2008).

Edwards, D., **Introduction to Graphical Modelling**, Springer (2000).

Faraway, J.J., Linear Models with R, Chapman & Hall/CRC (2005).

Feinberg, S.E., *The Analysis of Cross-Classified Categorical Data*, The MIT Press (1980)

Friendly, M., "Mosaic Displays for Multi-Way Contingency Tables," *Journal of the American Statistical Association*, 1994, Volume 89, No. 425, pp.190-200.

Hanneman, R. and Riddle, M, *Introduction to social network methods (2005).* On-line text available from: http://faculty.ucr.edu/~hanneman/nettext/Introduction_to_Social_Network_Methods.pdf

Hojsgaard, S., Edwards, D. and Lauritzen, S., *Graphical Models with R*, Springer (2012).

Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis*, Pearson (2007).

Kantardzic, M., *Data Mining*, Wiley (2011)

Kolaczyk, E.D., *Statistical Analysis of Network Data*, Springer (2009).

Lauritzen, S.L., *Graphical Models*, Oxford Science Publications (1996).

Mayer-Schonberger, V.and Cukier, K. *BIG DATA: A Revolution That Will Transform How We Live, Work, And Think.* Houghton Mifflin Harcourt, (2013).

Tan P.N., Steinbach, M. and Kumar V. *Introduction To Data Mining,* Addison Wesley (2006).

Whittaker, J., *Graphical Models in Applied Multivariate Statistics*, Wiley (1990).