008-Example of `knitr` and Large Documents

Car accident data

May 28, 2015

Abstract

We are interested in studying the relationship between car size and accident injuries. We first analyze a dataset containing counts of car accidents classified by accident type, severity of accident, car weight and whether or not the driver was ejected from the car, each of which have two levels of classification. A second table from the same study, which contains more detailed information on each of these categories, is further analyzed. We model the data using Poisson, logistic and multinomial logistic regression. Results find an increased risk of severe accidents among those who drive standard size cars (OR:1.40[1.2,1.7]), rollover(5.15[4.4,6.1]) and are ejected from the car (2.8[2.3,3.4]). Contrary to what previous data has shown, standard size cars had a higher proportion of severe accidents compared to small sized cars. Our study shows that the type of accident is a strong indicator of the severity of accidents.

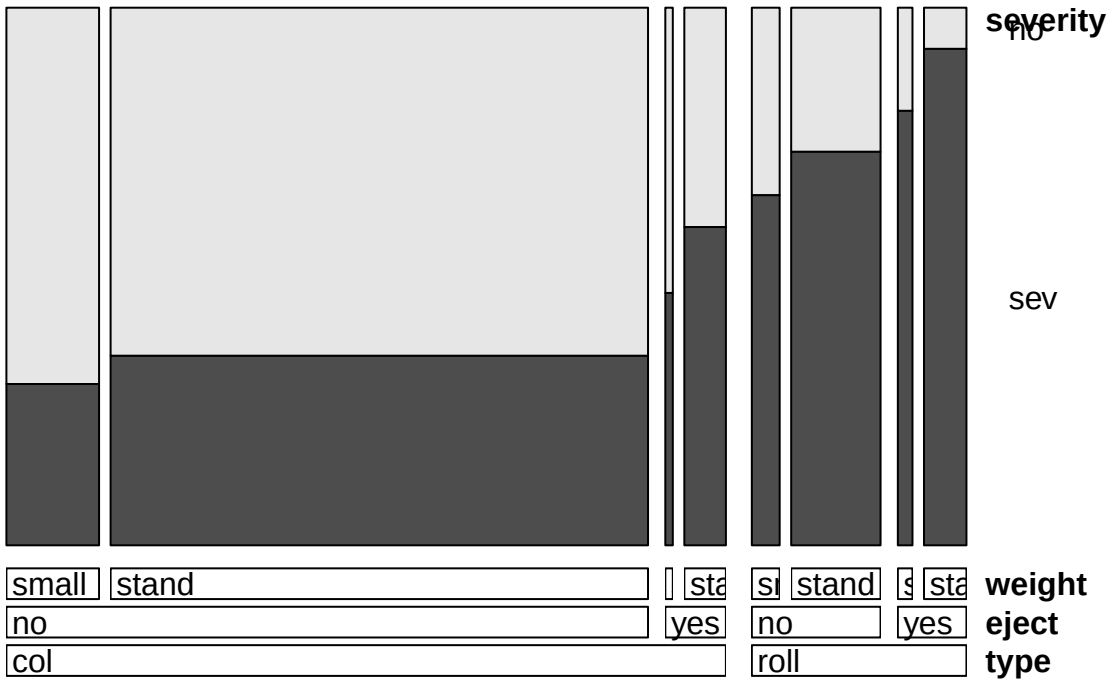
1 Introduction

Car size and weight are crucial to protecting people in crashes. Driver deaths in the U.S have been shown to decline fairly consistently as vehicle size increases Insurance Institute for Highway Safety (2009). Among 1 to 3 year old cars registered in 2007, the overall death rate in mid-size cars is 23% lower than in minicars. Due to growing environmental concerns in the past decade however, more automakers are reducing car weights to improve overall fuel economy. Environmentally conscious consumers are looking for smaller cars to reduce their ecological footprint. Increasing oil prices are another reason why market demand is shifting towards smaller, more fuel efficient vehicles. These statistics lead us to believe that we are safer in larger sizes cars which is a dilemma for the consumer as they try and get the best of both worlds i.e. a large enough car thats safe yet fuel efficient and good for the environment. We are interested in determining if there is a more complex relationship between the severity of accidents and car size. For example, if there are other factors such as the type of accident that can help predict accident severity. In section 2 we give a description of the datasets analyzed along with some plots describing our initial exploration. In section 3 we conduct a formal analysis using log linear and logistic models, followed by a discussion and conclusion in sections 4 and 5.

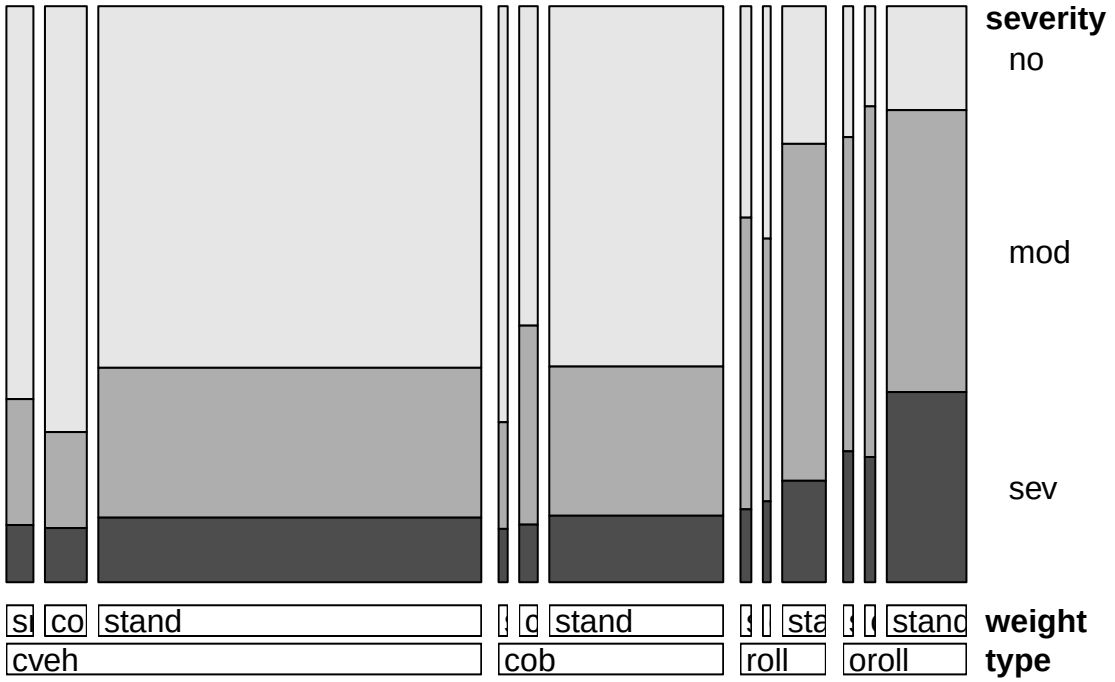
2 Description of Data

Information was obtained on 4,831 car accidents. The first dataset consists of four binary variables: the `weight` of the car involved in the accident (levels: `small, standard`), whether or not the driver was `ejected` (`no, yes`), the `severity` of the collision (`no, severe`), and the accident `type` (`collision, rollover`). The second dataset consisted of the same four variables, however three of them were polytomous instead of binary: `weight` (`small, compact, standard`), `severity` (`no, moderately severe, severe`) and `type` (`collision vehicle, collision object, rollover without collision, other rollover`).

Since we are primarily interested in factors affecting severity of car injuries, we construct a doubledecker plot (figure 1) which is constructed as a mosaic plot showing the conditional distribution of `severity` while simultaneously controlling for `type, eject, weight`. This plot shows the influence of the explanatory variables on `severity`, with severity being highlighted in dark grey. In figure 1a there is a very clear pattern i.e. the highest proportion of severe accidents is occurring in rollovers that were ejected from a standard car and then severity decreases all the way down to if the person was not ejected from a small car involved in a collision. We see a similar trend of the doubledecker plot of the more detailed dataset (figure 1b), the only difference being that some of the severe accidents have been re-classified as moderately severe as noted earlier. We note that a higher proportion of moderately severe and severe accidents occur in rollovers than in collisions.



(a) Simple Version of Data



(b) Detailed Version

Figure 1: Doubledecker plot for the Car Accident data showing the conditional distribution of car weight, given ejected, given type, and with severity highlighted

3 Analysis

For each of the datasets we fitted two models. For the simple dataset we fit a log-linear model and a logistic regression model with binary `severity`. For the more detailed dataset we again fit a log-linear model as well as a multinomial logistic regression to model `severity` which now has three response levels.

3.1 Simple Dataset

3.1.1 Log-Linear Model

We are interested in the association among the four variables, accident type (W), severity of accident (X), car weight (Y) and whether or not the driver was ejected (Z). The saturated log-linear model which fits the data perfectly, is given by (1)

$$\log \mu_{ijkl} = \gamma + \gamma_i^W + \gamma_j^X + \gamma_k^Y + \gamma_l^Z + \gamma_{ij}^{WX} + \gamma_{ik}^{WY} + \gamma_{il}^{WZ} + \gamma_{jk}^{YZ} + \gamma_{jl}^{XZ} + \gamma_{kl}^{YZ} + \gamma_{ijk}^{WXY} + \gamma_{ijl}^{WXZ} + \gamma_{ikl}^{WYZ} + \gamma_{jkl}^{XYZ} + \gamma_{ijkl}^{WXYZ} \quad (1)$$

We seek to find a simpler representation of the relationship between the 4 variables. To do this, we first fit the saturated model, and then compare it to a model without the 4th level interaction term using the likelihood ratio test (LRT). We continue in this manner by removing the 3rd and 2nd level interaction terms until we find the simplest model possible without significant loss of quality of fit. Table 1 summaries the four LRTs performed at $\alpha = 0.05$

Table 1: Likelihood Ratio Tests for Log linear models on Car Accident data

Model	Form	Residual Deviance	Residual d.f.	$P(\chi^2_{df} > \Delta D)$
0	(WXYZ)	0	0	NA
1	(WXY,WXZ,WYZ,XYZ)	0.67	1	0.41 (vs. Model 0)
2	(WX,WY,WZ,XY,XZ,YZ)	7.33	5	0.20 (vs. Model 1)
3	(W,X,Y,Z)	1193.11	11	0 (vs. Model 2)
4	(WX,WY,WZ,XY,XZ)	9.02	6	0.19 (vs. Model 2)

From table 1, we can see that there is no significant loss of information in the model when removing the 3rd and 4th level interaction terms, however there is a significant difference between the model with 2 level interaction and the model with no interaction. We further test to see if any of the 2nd order interactions can be dropped from the model and find that there is no interaction between car weight and whether the driver was ejected. We conclude that Model 4 provides the best fit to the data. This indicates these pairs of variables are conditionally independent i.e. they are associated in a pairwise fashion, but this degree of association does not depend on the level of the 3rd or 4th variable Wenyu Jiang (2013).

References

Insurance Institute for Highway Safety. Car size, weight and safety. https://www.iihs.org, 2009.

Wenyu Jiang. Log-linear models for 3-way contingency tables. https://www.mast.queensu.ca/~wjiang, 2013.

A Session Information

```
sessionInfo()
```

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 14.04.2 LTS
##
## locale:
##  [1] LC_CTYPE=en_CA.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_CA.UTF-8        LC_COLLATE=en_CA.UTF-8
##  [5] LC_MONETARY=en_CA.UTF-8    LC_MESSAGES=en_CA.UTF-8
##  [7] LC_PAPER=en_CA.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_CA.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils
## [6] datasets  base
##
## other attached packages:
## [1] vcd_1.3-2    knitr_1.10.5
##
## loaded via a namespace (and not attached):
## [1] colorspace_1.2-6 MASS_7.3-39      magrittr_1.5
## [4] formatR_1.2      tools_3.2.0      stringi_0.4-1
## [7] methods_3.2.0    stringr_1.0.0    evaluate_0.7
```