Analysis of Reported Drowsy Driving
Exploring Subpopulation Risk with Weighted Contingency Table Tools

Patrick Coyle*
and
Chen Chen, Nooreen Dabbish
Department of Statistical Science, Temple University

Abstract
The incidence of driving accidents due to human error, drowsy driving, in particular, is an important topic in the field of public health research, and might be considered preventable. Our study investigates the 606 self-reported drowsy driving accidents in 2013 as recorded in the General Estimates System (GES) from the Department of Transportation. This study seeks to understand the relationship between drowsy driving and conditional risk of fatality among different subpopulations and how this relationship changes depending on the time of day. We are exploring these interactions using recent developments in survey-weighted ROC analysis. By doing so, we hope to offer employers and government agencies insight into what can be done to reduce the rate of fatalities, especially those involving drowsy driving accidents.

Key Words: ROC, survey sampling, sleep, human factors, fatality, circadian behavior

1. Introduction

Driver sleepiness is an insidious and preventable cause of motor vehicle accidents, cited to under-reported [1]. A 1995 review of GES data found fatigue as a causal factor in 1.2 - 1.6% of crashes [2]. We extend and update this work to provide an understanding of the factors associated with drowsy driving. These include major factors cited in the literature to be associated with drowsy driving: commercial trucking, time of day and day of week, and road type.

We examine self-reported drowsy driving in the National Automotive Sampling System General Estimates System (NASS GES) data. This data represents a stratified sample of all police accident records and includes driver and passenger demographic information and accident condition and timing information.

2. Methods and Tools Used

In this study, we modeled whether or not a driver self-reported drowsiness as an impairment as the primary outcome. Univariate logistic regression for drowsy driving was used to predict and build an adjusted model for drowsy driving. Breslow-Day tests were used to compare staratifying effects between categorical predictors. Model comparison was carried out using ROC/concordance analysis.

All explored/modeled data is from the 2013 dataset; ROC analysis is conducted by fitting models to the 2013 dataset and using these models to predict drowsiness in the 2014 dataset. All programming and analysis was performed in `R`, using version 3.2.4 "very secure dishes." Thomas Lumley's `survey` package was used to implement survey methods.

*The authors gratefully acknowledge Dr. Richard Heiberger for his diligent supervision of this research.

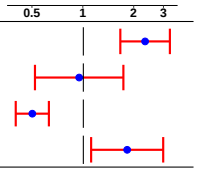
Variable	Odds Ratio	
Speeding	2.33	
Heavy Truck	0.94	
Female	0.50	
Interstate Highway	1.82	

Table 1: Univariate Testing of Categorical Predictors (Each factor has two levels. The odds ratio is the groups risk of drowsiness compared to its counterpart)

3. Analysis and Results

3.1 Predictors of Interest

Our primary goal is to construct and analyze a model to predict whether the police reported a given driver in an accident to be drowsy (in retrospect). The model's predictor variables were identified by i) reviewing prior research and ii) mass-univariate testing. Most variables required cleaning or transformation in order to maximize their predictive power. The predictors explored in this study are:

- Commercial Trucking: Binary variable, denoting Yes or No to whether commercial trucks are involved in the accident.
- Sex: Female or Male.
- Interstate Highways: Binary variable, denoting Yes or No to whether the accident happened on interstate highways.
- Relationship to Speeding: Binary variable, denoting Yes or No to whether whether the driver's speed was related to the crash as indicated by law enforcement.
- Age
- Hour of Day/Day of Week: 0 - 23, hours of each day, expanded on each day of the week.

3.2 Univariate Testing and Modeling

3.2.1 Modeling of Categorical Predictors

As shown in Table 1, *sex*, *incidence on an interstate highway*, and *speed-related accidents* appear to be strong predictors for drowsiness.

- Drivers in speeding-related accidents are more than two times more likely to be reported as drowsy compared to non-speeding drivers.
- There is no evidence from 2013 to indicate that a heavy-truck driver is more or less likely to be reported as drowsy than other drivers.
- Female drivers in accidents are half as likely to be reported as drowsy compared to males.
- Drivers in accidents on interstate highways are almost twice as likely to be reported drowsy compared to other sampled road types.

Effects	Strata	Trucker	Drowsy	Non-drowsy	OR	Pooled OR	P-value
Sex	Female	Y	1	47	2.759	1.272	< 0.001
		N	180	23,341			
	Male	Y	13	1,547	0.605		
		N	4127	29,669			
Age	≤ 24	Y	13	1,537	0.914	1.272	0.942
		N	381	41,152			
	> 24	Y	1	57	0.986		
		N	211	11,858			
Time of Day	7 am to 11 pm	Y	3	246	0.261	1.272	0.022
		N	223	4,777			
	12 am to 6 am	Y	11	1,348	1.067		
		N	369	48,233			

Table 2: Breslow-Day Test on Stratifying Effects within Trucker Population (Un-weighted Table)

3.2.2 Stratified Analysis of Drowsy Trucking

Truckers are a subpopulation that has received particular attention and intervention to reduce drowsy driving rates. We explore this population further by using other categorical variables (*sex*, *age* and *period of the day*) as strata and comparing the rates of drowsy trucking between strata using the **Breslow-Day test**. [1] The results are shown in Table 2, and can be summarized as follows:

- The risk of drowsy trucking is found to be much higher among female truckers than male truckers. However, there was only one drowsy female trucker sampled in 2013, which makes this result specious.
- There is no evidence that the risk of drowsy trucking is different for those under the age of 24 vs those over the age of 24.
- The risk of drowsy trucking is significantly higher between the hours of 12 AM and 6 AM.

2013 is a particularly interesting (and problematic) year to study drowsy trucking: between 7/1/2013 and 12/16/2014, the US Department of Transportation imposed a "restart" limit of 60 or 70 hours within a commercial trucker's "duty cycle," between which there must be at least two nighttime periods (from 1 A.M. until 5 A.M.). Therefore, we can stratify the data from 2013 by examining trucker data before this regulation (1/1/2013-7/1/2013) and after (7/1/2013-1/1/2014). The results, in the form of weighted mosaics, can be found in Figure 3.2.2. We find that the time-of-day distribution of drowsy truckers appears to be significantly changed after regulation, with accidents more likely to occur in the early morning. However, due to relatively low cell counts, this difference is not statistically significant (p = 0.212). There is also no evidence that the regulation reduced the total number of drowsy truckers in the second half of the year compared to the first.

A matter of concern is the low count of total drowsy truckers sampled in 2013; pooling additional years of data could strengthen this result.

[1] There is currently no documented variance estimation for a survey-weighted Breslow-Day test. Instead, we test using the raw table counts instead of the sum of weights.

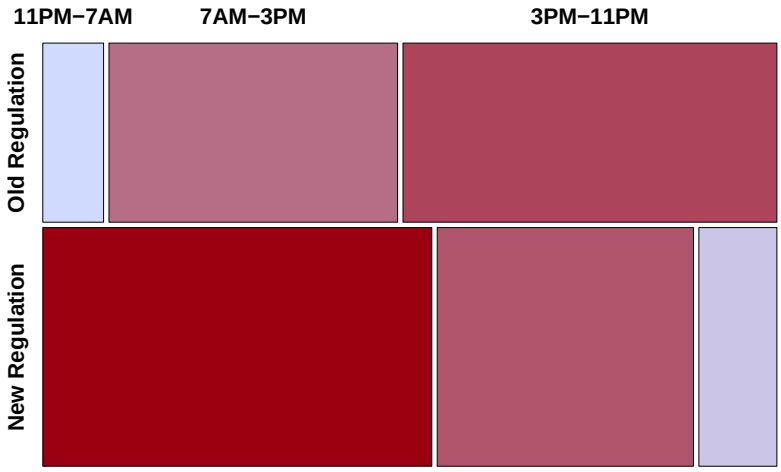
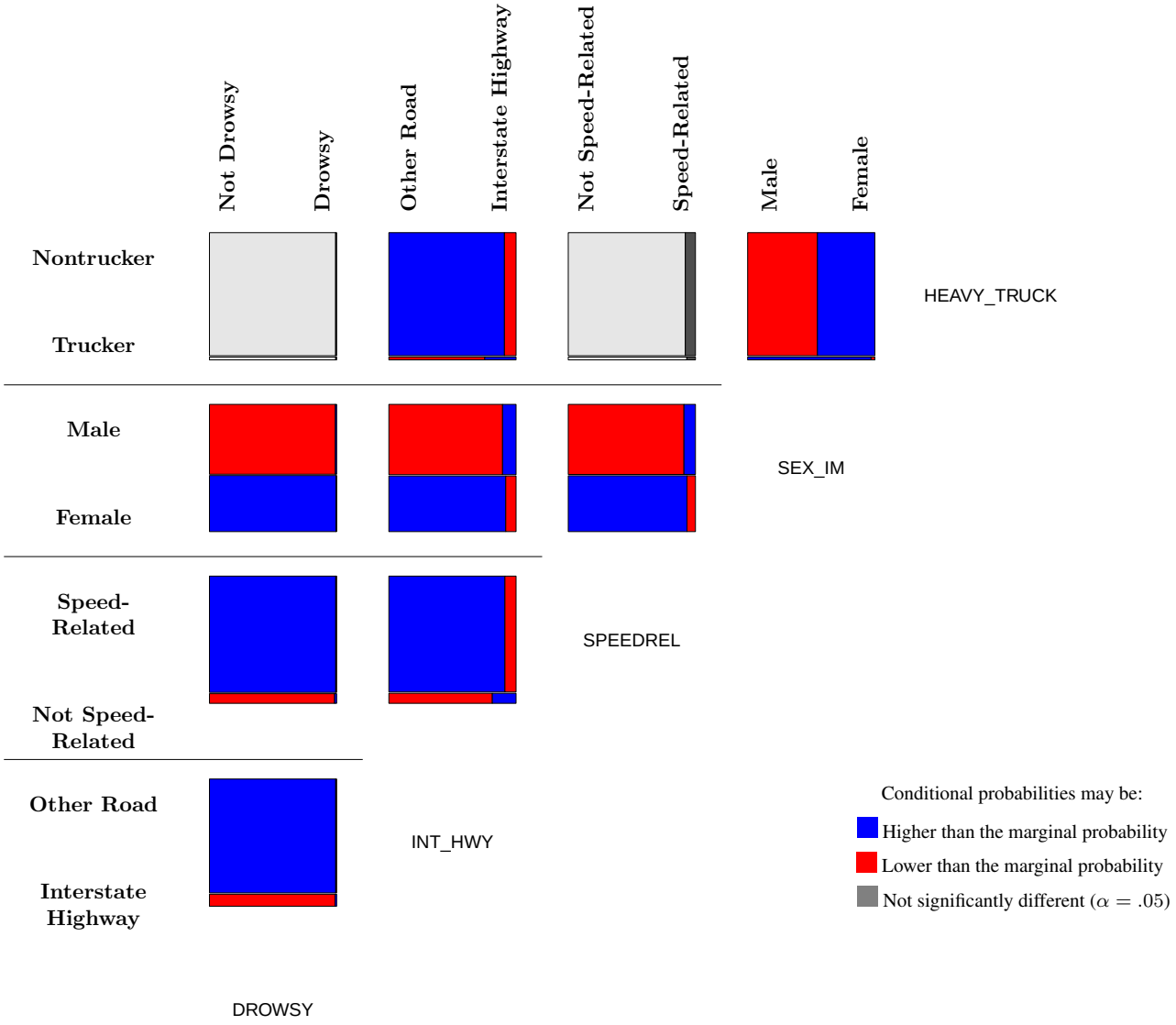


Figure 1: Weighted mosaic for the time-of-day distribution of drowsy truckers before and after regulation on 7/1/2013

We also explored the correlation between categorical variables, as shown in a ”mosaic matrix” in Figure 3.2.2. Mosaics are analogous to scatterplots and are a good alternative when dealing with categorical or discrete data. We color these mosaics using the results from survey-weighted hypothesis tests in order to gain valuable insights. For example:

- The mosaic in cell (1, 3) of the matrix reveals that there is no significant difference in speeding-related accidents between non-truckers and truckers.
- Males dominate the commercial trucking industry, which is reflected by the mosaic in cell (1, 4) of the matrix.



1

Figure 2: Weighted Mosaic Matrix: Interaction of Predictors

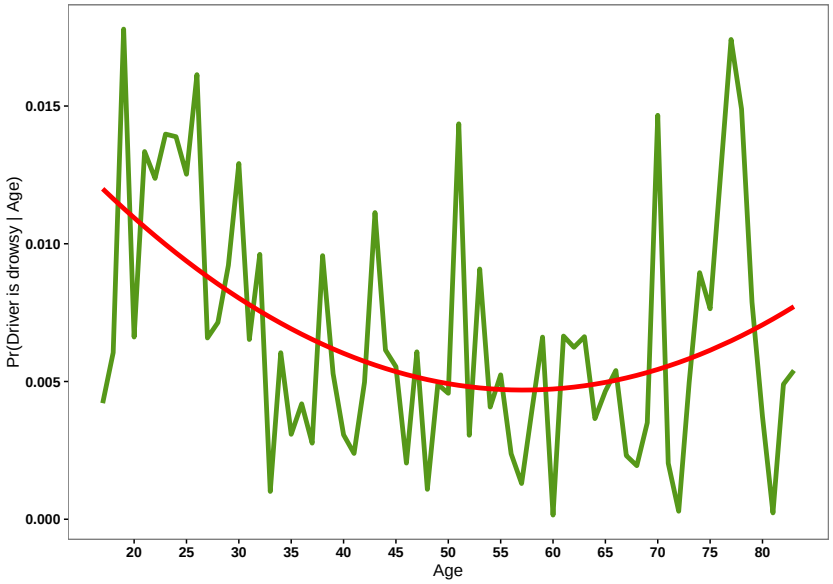


Figure 3: Age vs. Drowsy Driving (Age is a very noisy predictor for risk of drowsiness, but it does appear to have a weak trend that could contribute to prediction.)

3.2.3 Modeling of Continuous Predictors

Age and *hour of day* are two continuous factors that may have strong predictive power for drowsy driving. We observed each factor and their relationship with probability of drowsy driving. The results are shown in Figure 3.2.3 and Figure 3.2.3.

Age: The relationship between age and drowsiness was discussed in above section, where age was dichotomized into two groups, ≤ 24 and > 24. Observed from unweighted data, younger age group seems to have same chance of having an drowsy accident compared to older age group. This result runs counter to prior research for the general population, which indicates that younger drivers have a higher risk of drowsy driving. We explore this relationship further using the raw, non-dichotomized age data for the general population. Shown in Figure 3.2.3, age appears to be a noisy predictor for the risk of drowsiness, but there is weak quadratic trend which may contribute to the prediction, so we fit this transformation to the model.

Hour of Week: This is a combination of two variables, Hour of Day (HOUR_IM) and Day of Week (WKDY_IM). It ranges from 0 to 168, with 0 being 12 a.m. on Sundays and 168 being 11 p.m. on Saturdays. The relationship between this variable and risk of drowsiness yields two trends:

- A cyclic trend within a day – the risk picks up before midnight, with a peak in the 12 a.m. to 8 a.m. interval, and small jumps but mostly static risk level for the rest of the day.
- A slight cyclic trend at the day-of-week level, with peaks on the weekend and a nadir at midweek.

We fit the following transformation into the model, where t is the time of week (ranging

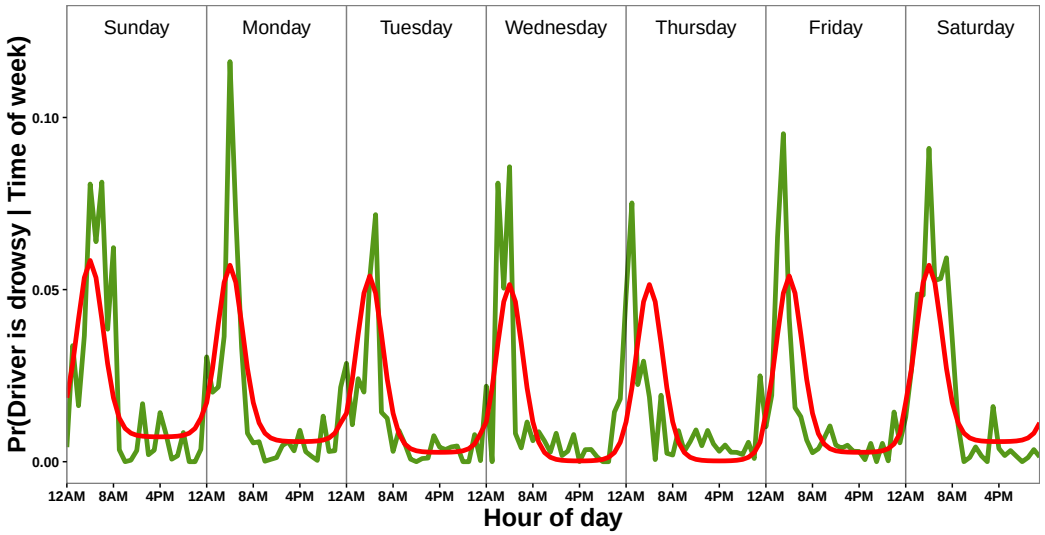


Figure 4: Hour of Day/Day of Week vs. Drowsy Driving (Time is a very strong predictor of drowsiness, especially at the hour-of-day level)

from 0 to 167):

$$y = .0026e^{3\cos\left(\frac{2\pi(t-4)}{24}\right)} + .0037\cos\left(\frac{2\pi\left\lfloor\frac{t}{24}\right\rfloor}{7}\right) + .0034$$

The transformed variable fitting y is shown in Figure 3.2.3 .

3.3 Model Construction and Diagnostics

We used the above variables to fit various models to predict the risk of drowsiness for each sampled driver in 2014, the subsequent year of accident data. We then evaluate the quality of these predictions by calculating the models concordance, comparing each true positive (drowsy) observation against each true negative (non-drowsy) observation in a pairwise manner.

The *concordance statistic* is defined as the percentage of true-positive/true-negative pairs for which the true positives' predicted risk is greater than the true negatives' predicted risk, over all pairs of predicted values. In other words, if a driver is reported to be drowsy, s/he should have higher predicted risk than the driver that is not reported to be drowsy, in order to contribute to the nominator concordance statistic.

A value of 0.5 indicates the model is no better than predicting using random chance, and if the statistic falls below 0.5, it shows that the model is very poor. We have plotted the concordance using the Receiver Operating Characteristic Curve (ROC) adjusted to complex survey weight, with understanding that, in the case a binary response (as we have here), the area under the ROC curve (AUC) is equivalent to the concordance statistic. ROC curves for three separate models are shown in Figure 7. These are calculated using the `WeightedROC` package, which may be found on Github.

Before including the transformed variables for *age* and *hour of week*, we could reach a concordance score of 0.61 (categorical predictors with interactions). After including *age* and *hour of week*, we boosted the prediction power much more, with a concordance score of

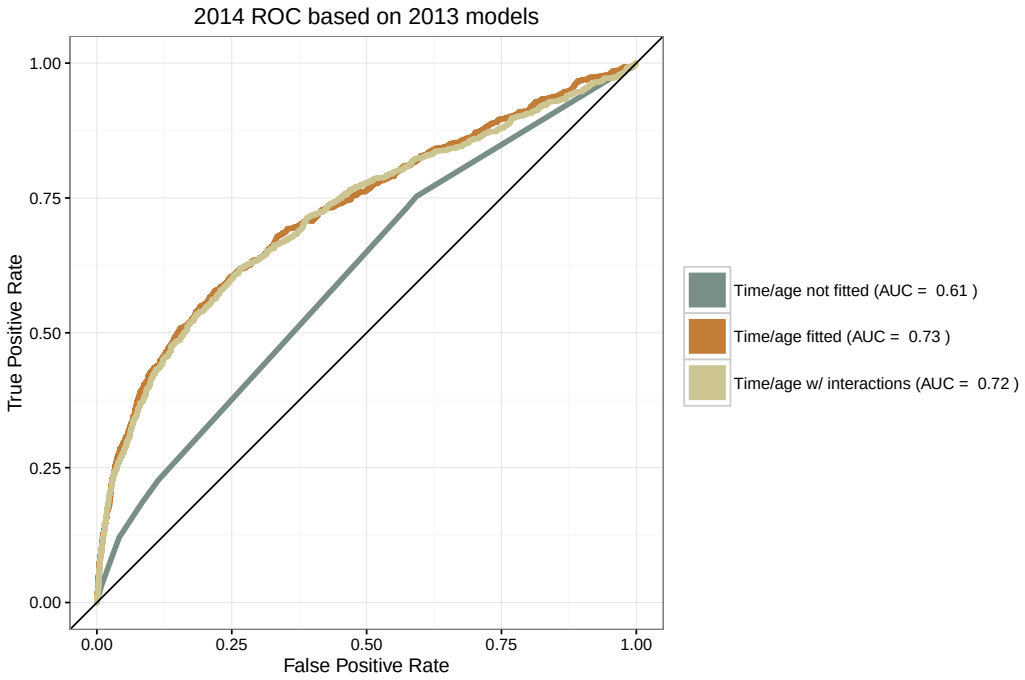


Figure 5: ROC curves of fitted 2014 data using 2013 model

0.73. We find that including interactions between the categorical and continuous predictors slightly lowers our concordance score, presumably due to overfitting.

4. Conclusion, Limitation and Future Works

The goals for this study were to investigate the self-reported drowsy driving accidents, to understand the risk of accidents related to drowsy driving among different sub-populations and how it changes depending on the time of day, then to explore these interactions using recent developments in survey-weighted ROC analysis.

By using complex survey tools built in Thomas Lumley's `survey` package in `R`, and utilizing survey-weighted data visualization tools, we were able to understand that *hour of week, age, sex, speeding*, and *interstate highways* are important predictors for drowsy driving. By using the sample-weighted ROC analysis, we showed that we could obtain a concordance score as high as 0.72.

The original calculation of concordance involves matching pairs. We might decrease this statistic's bias by weighting the C-statistic using joint inclusion probabilities of the pairs (Yao et al, 2013). These are not strictly the product of two inverse weights; they depends on where the paired observations belong in relation to one another in the stratified sampling hierarchy (sampling without replacement). As a limitation to this study, the data only has the total weights, which is the product of the probabilities at each of the three levels of the survey design. We were only able to successfully petition for the PAR-level weights from the NHTSA due to privacy concerns. We must settle for using individual weights. Other survey-weighted research settings may benefit from exploring and programming this theoretical result.

We could learn much more about the demography of drowsy driving by fusing GES data with the American Community Survey (ACS), exploring alcohol/drug use as a confounding component, and exploring bias in the reporting of drowsy driving.

As discussed in our stratified analysis, the `survey` package could be meaningfully improved by formulating a variance estimator for survey-weighted Breslow-Day testing.

Appendix

R-package for analyzeGES routine: R-package `analyzeGES` containing code to perform the diagnostic methods described in the article. The package also contains all datasets used as examples in the article. The structure of the package is as follows:

- R functions
  - `mergeGivenYear(dir, desired.files)`: Allows user to specify a list of GES data file names from a desired year's directory, then reads and merges the named files from the .sas7bdat format in which the NHTSA stores them.
  - `mosaicSplom2(response = "DROWSY", predictors, ...)`: Produces an upper triangular matrix a two-way weighted mosaic plots for the input set of variables. This function currently requires that the response and predictors are all factors. This is written as an alternative to the scatterplot matrix for situations in which all variables are factors. This is written out of desire to display the data with mosaics as opposed to other common discrete-discrete plotting defaults, such as faceted bar charts or jittered scatterplots, which do not precisely communicate the full pairwise joint distributions.
  - `predROC(glm.obj, newData, ...)`: Predicts the response/label for a test set based on the svyglm object created from the test set and calculates the ROC curve for the predictions.
  - `model_and_convert(response, predictors)`: Creates a list of three models (one with just the predictors and response, and two with varying levels of interaction with transformed time and age predictors, as described in the article), applies `predROC` to each item in the list, and formats the outputs into a long data frame for convenient visualization with `ggplot`.
  - `GES_plotter()`: Launch a Shiny application that allows you to choose a response, set of predictors and desired plot type (mosaic matrix, GGPairs or ROC curves). This shiny app is also hosted at https://patrickcoyle.shinyapps.io/GES_plotter/
- Datasets (for 2013 and 2014, as explored in this paper)
- Data-cleaning and variable-transformation script

This package is hosted on Github and can be downloaded using `install_github("PatrickCoyle/analyzeGES")`.

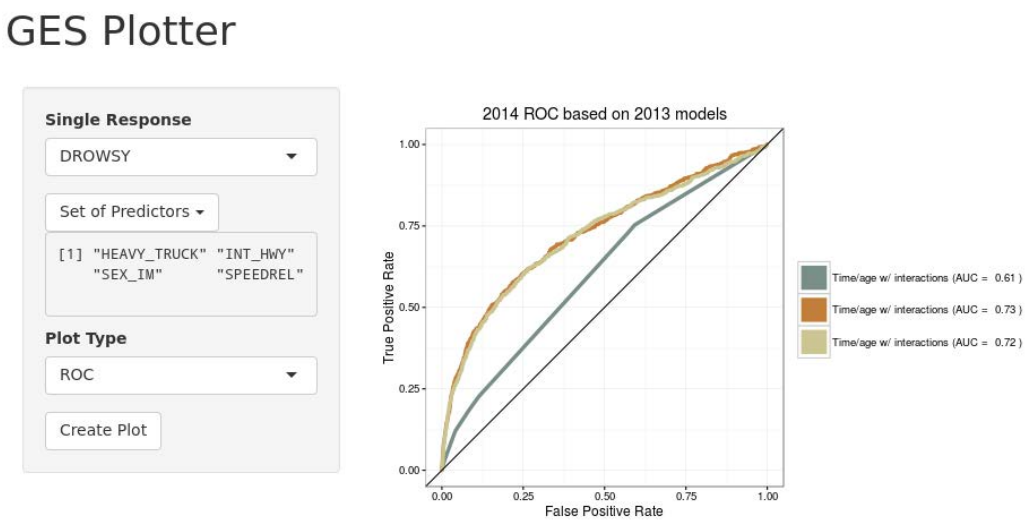


Figure 6: GES_plotter, a Shiny application for visualization of survey-weighted models and their ROC performance. This allows for convenient exploration of the dataset by automating the project's model-building and visualization process with a graphical user interface. With more work to define additional input options, this application will be portable to other research projects involving weighted or unweighted data with a binary response.

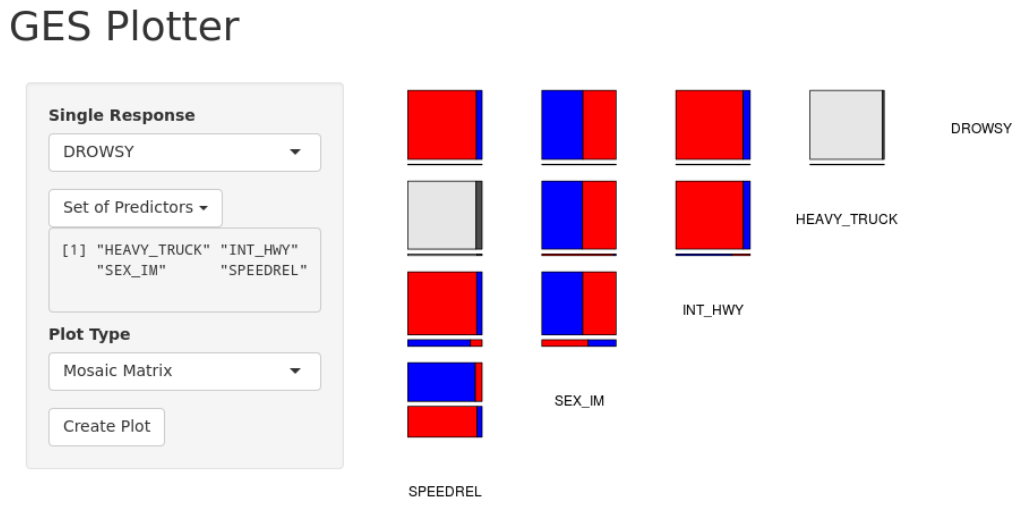


Figure 7: GES_plotter used to print a matrix of weighted mosaic plots for a user-input subset of variables.

References

[1] Watling, Christopher N., and Hanna Watling. "Sleepy driving and drink driving: attitudes, behaviours, and perceived legitimacy of enforcement of younger and older drivers." (2015).

[2] Knipling, Ronald R., and Jing-Shiarn Wang. "Revised estimates of the US drowsy driver crash problem size based on general estimates system case reviews." Annual

proceedings of the Association for the Advancement of Automotive Medicine. Vol. 39. Association for the Advancement of Automotive Medicine, 1995.

[3] Austin, Peter C. and Ewout W. Steyerberg. ”Intepreting the concordance statistic of a logistic regression model: relation to the variance and odds ratio of a continuous explanatory variable.” BMC medical research methodology, 12(1):82-82, 2012.

[4] Lumley, Thomas and Wiley-Blackwell Online Books. ”Complex surveys: a guide to analysis using R.” John Wiley, Hoboken, NJ. 1st edition, 2010.

[5] Yao, Wenliang, Zhaohai Li and Barry Graubard. ”Estimation of ROC curve with complex survey data.” Statistics in Medicine, 34(8):1293-1303, 2015