Final Project

Nate Koser

Rutgers University

Final Project

Introduction

This paper is an investigation of drunk driving patterns in South Korea for the year 2016. The data were obtained from kosis.kr, the Korean Statistical Information Service website maintained by the National Statistical Office. The data include information on the number of cases, deaths, and injuries resulting from drunk driving incidents organized by time of day, month of the year, and level of intoxication.

Hypotheses

The analysis presented here is an investigation of three main hypotheses. The first is that the number of cases that occured will be a good predictor for time of day (day or night), and that generally there are more incidents at night. The second is that there will be more cases in months with major holidays. These months will include December/January - when solar new year's day occurs; February - when lunar new year's day occurs; and September - when the fall harvest holiday (similar to Thanksgiving in North America) occurs. The third hypothesis to be tested involves high levels of intoxication. Within the `abv` variable, there is a value ">.35%", indicating a blood alcohol content of over .35%. The hypothesis is that at this extreme level of intoxication, the risk of death will be greater. Descriptive statistics, plots, and model information for all three hypotheses will be provided in the following sections.

Data analysis

Day and Night

The following table shows a summary of some descriptive statistics regarding the general occurance of drunk driving during the night and during the day. The numbers given are log-adjusted.

timeofday	CasesMean	CasesSd	InjuriesMean	InjuriesSd
day	2.92	1.63	3.40	1.73
night	3.58	2.01	4.05	2.09

We observe a higher mean incidence of cases during the night than during the day, as well as a higher mean number of injuries during the night vs. the day. This is line with our proposed hypothesis.

The following is a plot of a binomial regression where a value of 1 indicates "day" and a value of 0 indicates "night". The predictor is the number of cases that occured.

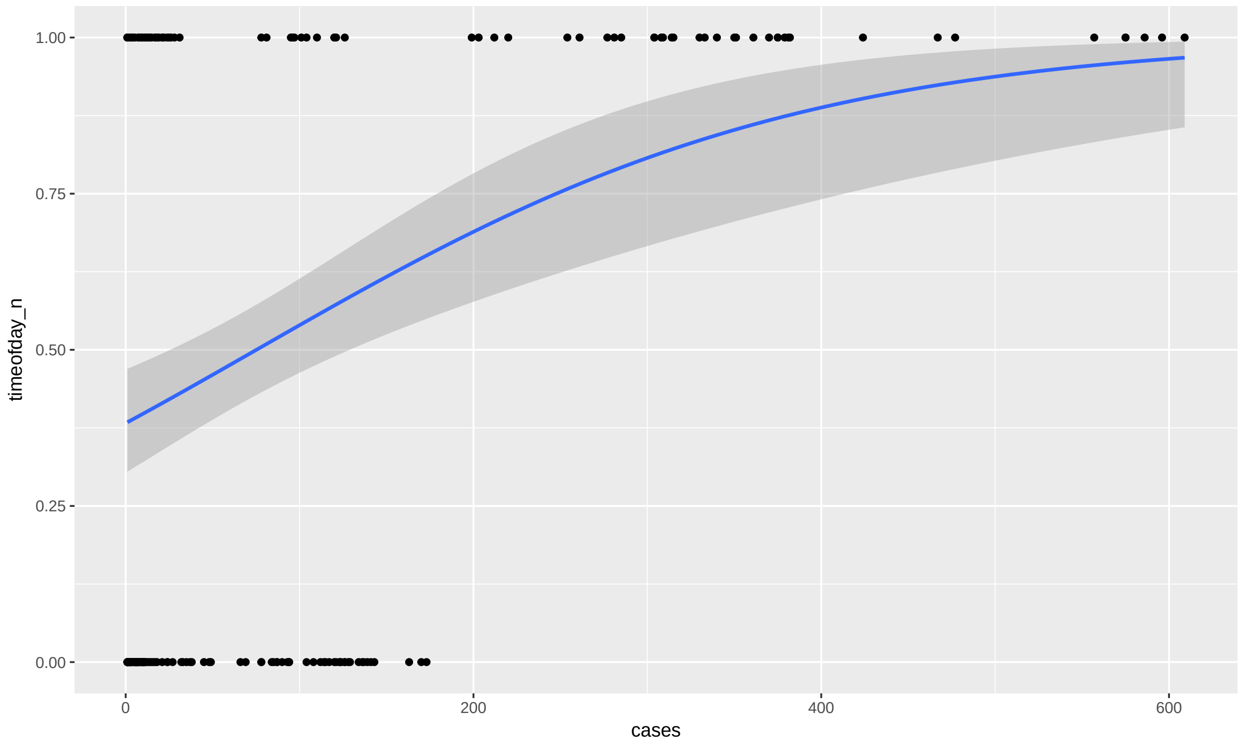

Figure 1

Visual inspection reveals a general upward trend towards nighttime as the number of cases increases. In fact, the max number of cases occuring in any one daytime period (~180) appears to be much lower than the max number of cases occuring on the busiest night (>600). The shape of the curve reflects this and it is once again amenable to the stated hypothesis.

A generalized linear model was fit to the data using a binomial regression with a logit linking function. The following table is a summary of the results of this model fit.

term	estimate	std. error	z value	p value
(Intercept)	-0.48	0.18	-2.66	< 0.01 *
cases	0.01	0.001	4.25	< 0.001 **

We see that there is a significant effect of cases as a predictor for night or day in the data (*p*-value <0.001). Taking the inverse logit of the estimates for the night-day intercept and cases reveals a 62% chance of the incident having occured at night when the number of cases equals 100, and an increase in this chance by ~20% when the number of cases is 200. Increasing the number of cases to 300 reflects a further ~10% increase in the probability of the event occuring at night, up to 92.55%. These results are reflective of the trend observed in the plot above.

Months of the Year

The following table shows a summary of some descriptive statistics regarding the general occurance of drunk driving during different months of the year. The numbers given are log-adjusted.

month	CasesMean	CasesSd	InjuriesMean	InjuriesSd
April	3.43	1.93	4.00	1.83
August	3.26	1.86	3.68	2.02
December	3.30	2.05	3.68	2.25
February	3.16	1.99	3.78	1.90
January	3.22	1.93	3.61	2.13
July	3.06	2.01	3.55	2.02
June	3.10	1.80	3.55	1.95
March	3.31	1.94	3.82	1.96

month	CasesMean	CasesSd	InjuriesMean	InjuriesSd
May	3.38	1.80	3.93	1.79
November	3.36	2.01	3.74	2.20
October	3.34	1.87	3.81	2.00
September	3.25	1.75	3.74	1.88

The months described in the hypothesis above do not appear to have any special status in terms of the number of cases or injuries that occur. In particular, Janurary and February's cases and inuries are among the lowest, while April sticks out as being the most active. Also of note is the standard deviation values, which are more than 50% of the corresponding mean for most values in the table. This indicates data that is more spread out, and less concentrated around any one point (the mean).

Figure 2 is a boxplot of the log cases of drunk driving with the month as the predictor.

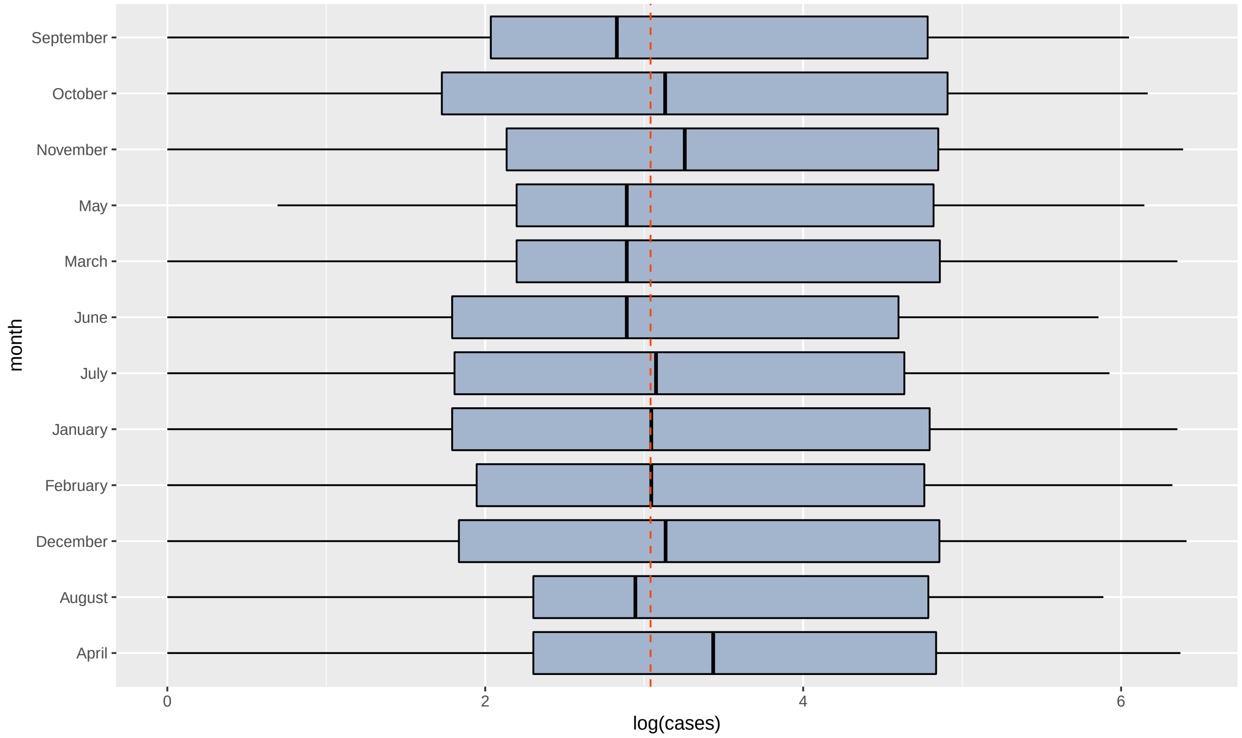

Figure 2

The median of the log cases for all months is 3.04, and the vertical red line in the plot marks this value. Visual inspection reveals that the median of the log case value for January and February is right on the the red line. The value for December is slightly above it, and the value for September appears to be lower. This is not in line with the proposed hypothesis. Unexpectedly, April shows the biggest departure from the median value - it is the highest above the red line.

As December is the month among our hypothesized months with the highest mean and median cases, it makes sense to isolate it as a “best case scenario” and fit models considering it. To this end, I fit a binomial generalized linear model to the data, where a response of 1 indicates that the event occured in December, and a response of 0 indicates that the even occured in some other month. The results are summarized in the table below.

term	estimate	std. error	z value	p value
(Intercept)	-2.56	0.32	-7.95	< 0.001**
cases	0.001	0.002	0.5	0.6

No significant effect of cases was found on an event occuring in December or not. An investigation of inverse logit values reveals an 8% chance of having occured in December when the number of cases is equal to 100. An increase to 200 cases comes with a positive change of <1% probability of being in December, and a further increase to 300 cases has the same result, raising the probability to only 9.4%. None of the methods employed here confirm our hypothesis that drunk driving incidents are more frequent in months with a major holiday.

High Blood Alcohol Content

The following table shows a summary of some descriptive statistics regarding drunk driving at a measured blood alcohol content of over and under .35% and the mean deaths per case for each category. The numbers given are log-adjusted.

BAC	Deaths as percent of cases
over .35	0.03
under .35	0.02

The numbers indicate that there is a slightly higher chance of death at the advanced BAC of >.35% than there is for lower BAC values.

The following is the plot of a binomial regression where a response of 1 indicates that the event occured at a BAC of >.35%, and a response of 0 indicates otherwise.

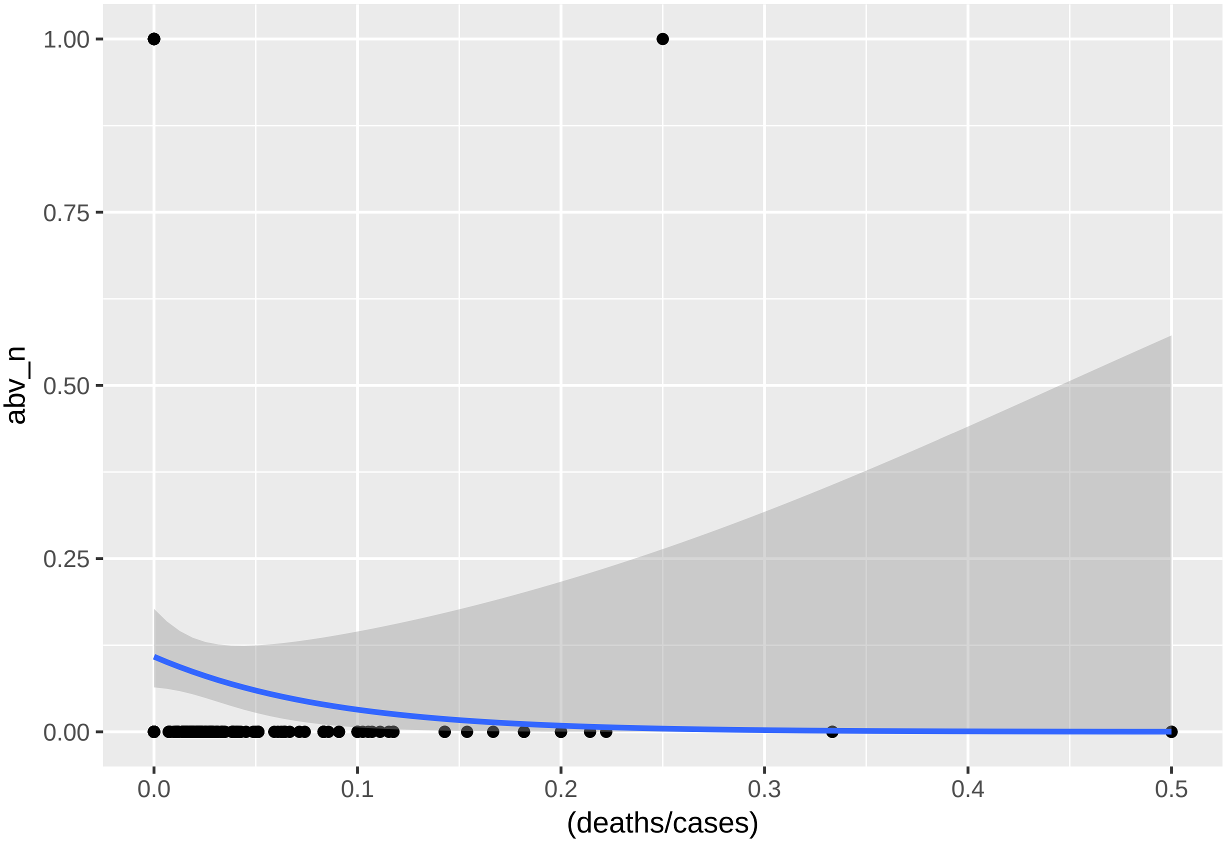

Figure 3

The plot indicates that there is a slight negative trend - as the number of deaths per cases increases, the chance that the incident in question occured at a BAC value of >.35% decreases. This is not what the descriptive statistics in the table above indicated. This may be due to the scarcity of cases occuring at such a high BAC - the larger number of cases where a larger propotion of people died at a lower BAC could offset the fewer number of cases at a higher BAC.

The following table is the result of fitting a generalized linear model with a binomial, logit linking function to the data.

term	estimate	std. error	z value	p value
(Intercept)	-2.11	0.29	-7.21	< 0.001**
deathchance	-13.04	9.37	-1.39	0.16

There is what could be considered a non-significant trend (p = .16) for chance of death when being invovled in an incident at a BAC of over .35%. This trend is negative, as can be observed in the plot. Taking the inverse logit indicates a 10.8% chance of the incident occuring at a BAC of over .35% when the number of deaths per case is equal to zero. When the deaths per case increases to .5, the chance of having occured at BAC >.35% is a mere .02%. This does not confirm the hypothesis that more deaths occur at a higher blood alcohol content.

Discussion

In the preceding analysis, three hypotheses were proposed and analyzed. The first hypothesis was that there are more cases of drunk driving during the night than during the day. Statistical analysis through generalized linear models revealed that this is the case - there was a significant effect of number of cases on time of occurance (p-value < 0.001).

The second hypothesis was that more drunk driving incidents occur in months that have major holidays. These included September, December, January, and February. Statistical anaylsis showed that there is no such effect. For the month of December, which was determined ot be the most hopeful in terms of the hypothesis, there was only a slight positive trend which was not found to be statistically significant (p-value < 0.6).

The third hypothesis was that at the high BAC level of >.35%, there would be a higher number of deaths per incident of drunk driving. While examination of the mean initially suggested that this may be the case, there was no statistically significant effect found in support of the hypothesis. In fact, a general negative trend was found (*p*-value .16). This non-intuitive result may be due to the dearth of cases at this high BAC - visual inspection of the data set reveals that they are only a tiny portion of the overall number of cases. It is possible that incorporation of more data would reverse this trend.