Supplementary Document for Traffic Fatality Project

Boyang Zhang

Data acquisition

As summarized in Data collection section in our report, we mainly obtained our raw data from three resources, namely Fatality Analysis Reporting System of the National Highway Traffic Safety Administration, U.S. Census Bureau and Local Area Unemployment Statistics. We saved our code for data downloading as "0-download.R". Aside from post 2010 county-level population estimates, all our data can be automatically downloaded by sourcing "0-download.R". Our data was obtained on October 1st, 2015 and saved them in raw_data file under data directory.

Variable definitions and Missing value imputation

In this section, we would mention about how we defined variables and imputed missing data for drunk driver variable.

FIPS county code

The FIPS county code is a five-digit Federal Information Processing Standard (FIPS) code (FIPS 6-4) which uniquely identifies counties and county equivalents in the United States [2]. In our analysis, we used county FIPS code as one of the merging columns to combine all datasets. Here, we incorporated "county.fips" dataset in "maps" r package as a source to match FIPS county codes and their county names [3].The following table provides current county FIPS code for all 24 counties in Maryland.

Table 1: An overview of 24 counties in Maryland and their FIPS code

FIPS	County
24001	allegany
24003	anne arundel
24005	baltimore
24009	calvert
24011	caroline
24013	carroll
24015	cecil
24017	charles
24019	dorchester
24021	frederick
24023	garrett
24025	harford
24027	howard
24029	kent
24031	montgomery
24033	prince georges
24035	queen annes
24037	st marys
24039	somerset
24041	talbot
24043	washington
24045	wicomico
24047	worcester

FIPS	County
24510	baltimore city

Accident files - light condition & weather

For accident level files from 2005 to 2015, we first extracted common columns shared by all accident files and then concatenated them together. According to multi-year analytical user's manual provided by Fatality Analysis Reporting System (FARS) [4], we recategorized light condition and weather as follows. For light condition, we categorized night with no light as one level, dawn,dusk and night with light as second level and daylight as the third level. That is to say, the larger number of light condition, the brighter or clearer view that a driver has. As for weather, we simplified the original categories by defining clear or adverse weather. We defined all non-reported or unknown cases as missing values in both variables.

After redefining categories, we summarized accident data by county and date (i.e. year and month). We extracted average light condition and calculated proportion of severe weather when accidents happened for a specific county in a given month.

All our preprocessing code for accident level files are saved in "1-acc.R" under R code directory.

Person files - age & drunk drivers

For person level data, we first extracted common columns from 2005 to 2015 and concatenated them. Then we included only drivers' information by specifying person type as 1 according to manual [4]. For drunk drivers variable, the data format changed in 2015 and therefore we multiplied alcohol test result before year 2015 by 10 to share the same scale. According to Maryland's DUI laws, drivers whose blood alcohol content is larger than 0.08 are defined as drunk driving. By applying the threshold of blood alcohol content at 0.08, we defined drunk driving based on alcohol test result (>0.08) and imputed the missingness of drunk driving action reported by the police. In this way, by summarizing data for each county for each month, we obtained average drivers' age that are died due to fatal crashes and proportion of drunk drivers involved in fatal crashes. All preprocessing code for person level files are saved as "1-per.R" under R code directory.

Population

Population density is also a factor related to traffic-related fatalities. It is reasonable to guess that higher risk for fatal crashes among denser population counties. The data we collected are from two resources and for post 2010 population we could only download by hand. By matching county names with their FIPS code described above, we combine pre-2010 and post-2010 population estimates by county and year. In our analysis, we assumed that population estimates remained same throughout a year. The code for preprocessing population files are saved in "1-population.R" under R code directory.

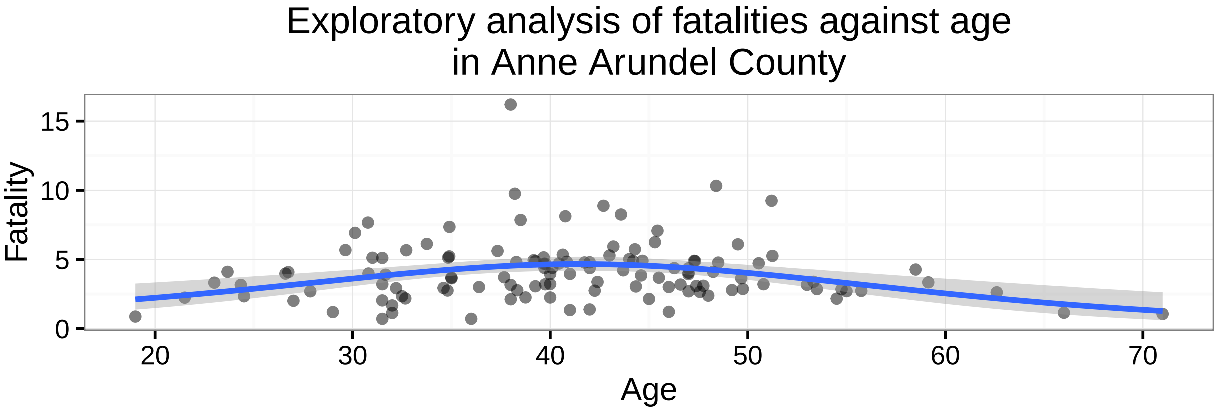
Macroeconomic conditions

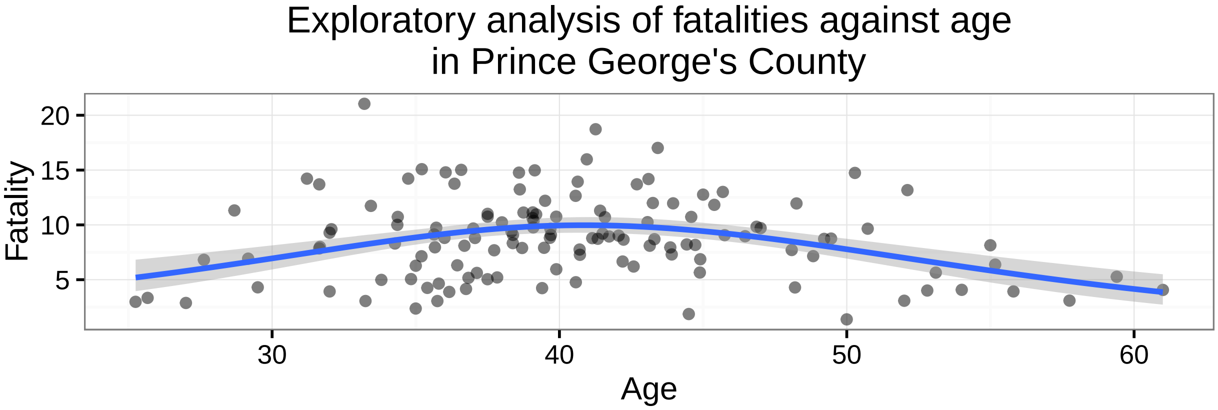
For macroeconomic conditions, we collected data for both median household income and unemployment rate. For median household income, we only obtained annual median household income for county level and therefore we chose to focus on unemployment rate instead. Similarly, by matching county names and their FIPS code, we combined unemployment rate from 2005 to 2015 based on county and date (i.e. year-month).

Exploratory analysis

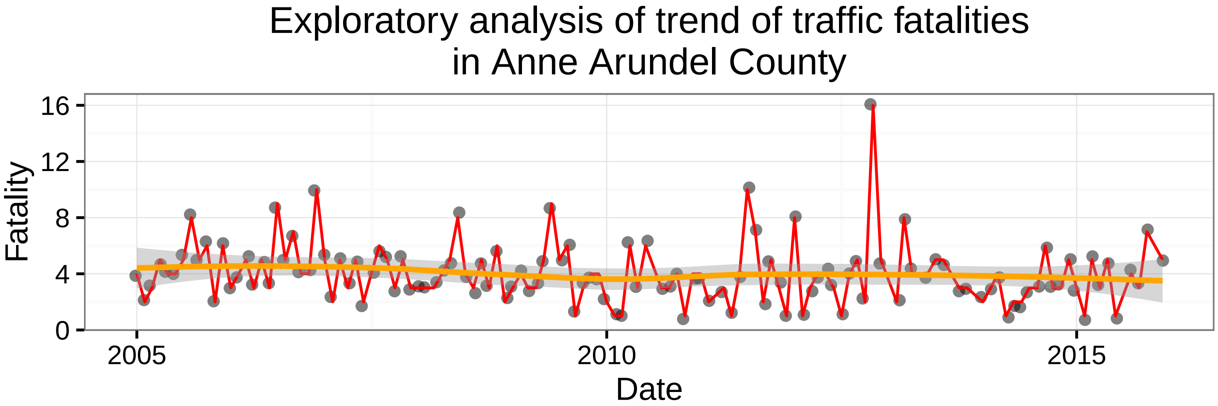
After defining variables, we performed a thorough exploratory analysis on each county. In report, we covered exploratory plots of Baltimore City. Here, we used Anne Arundel's County and Prince George's County as examples to illustrate results.

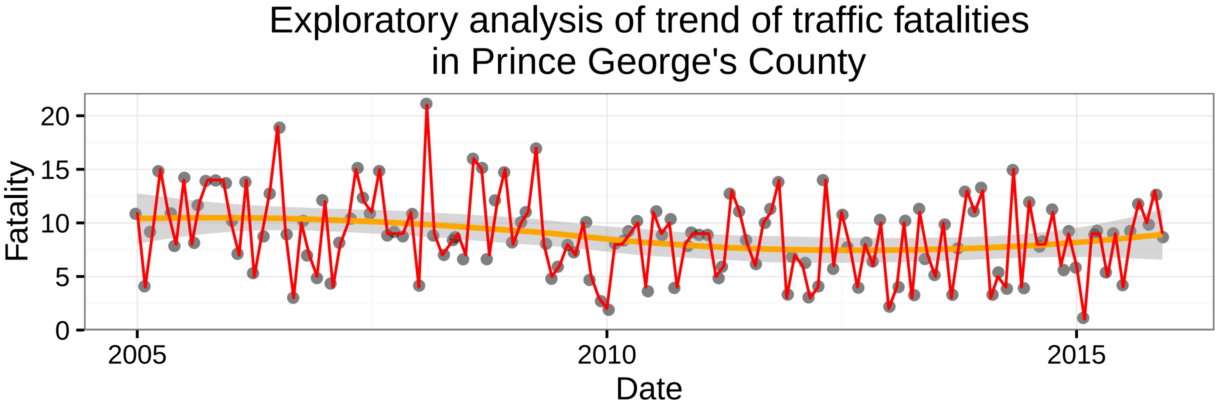
Firstly, we explored traffic-related fatalities against age. In the following two plots, the dot represents data while blue lines represents a smoothing line using natural spline with 2 degrees of freedom. The shaded area are point-wise confidence interval. We noticed that the fatalities displayed different patterns before 40 and after 40 years old. Hence, we applied a linear spline for age at 40 years old.



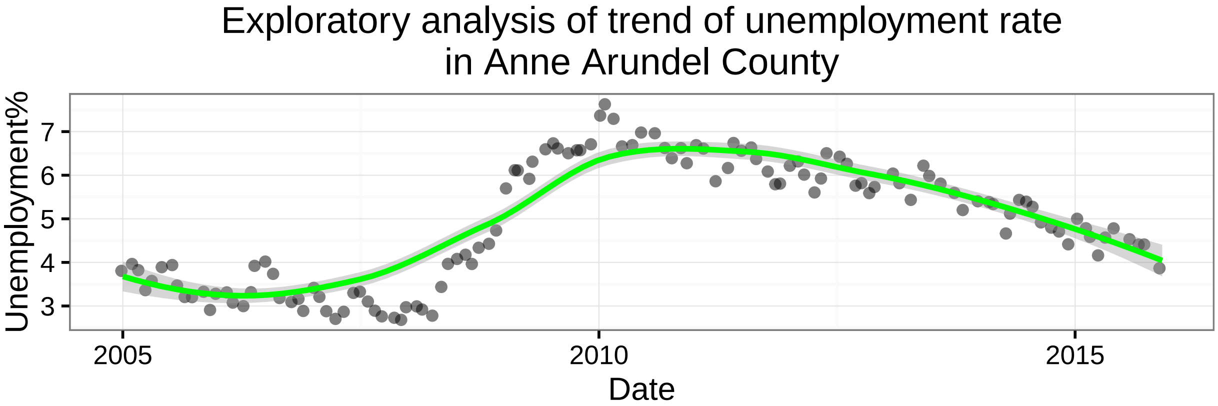


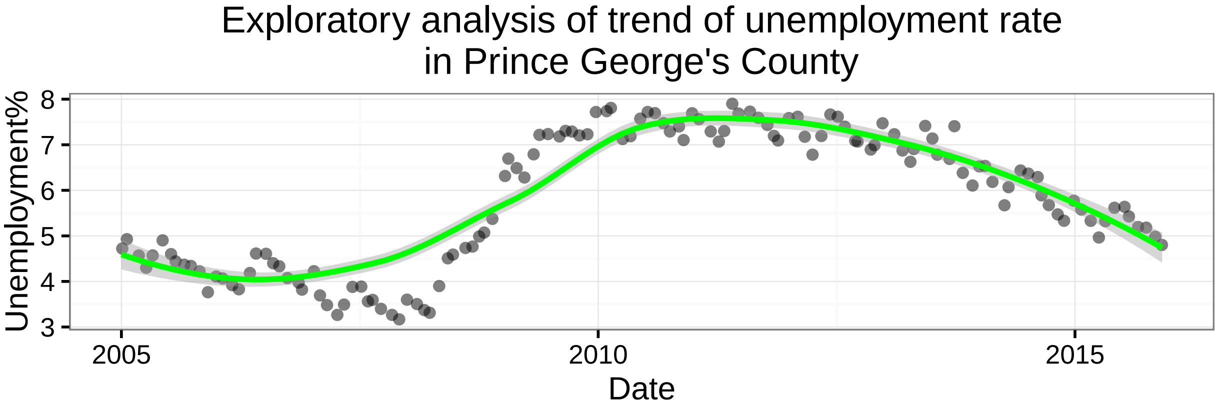
Secondly, we plotted trend of fatalities in each county. In the following two plots, the red line represents true fatalities and orange lines are loess smoothing with shaded area indicating point-wise confidence interval. Although different counties displayed different seasonality, we decided to use natural spline of date(i.e. year-month) with two degrees of freedom to account for seasonal patterns.



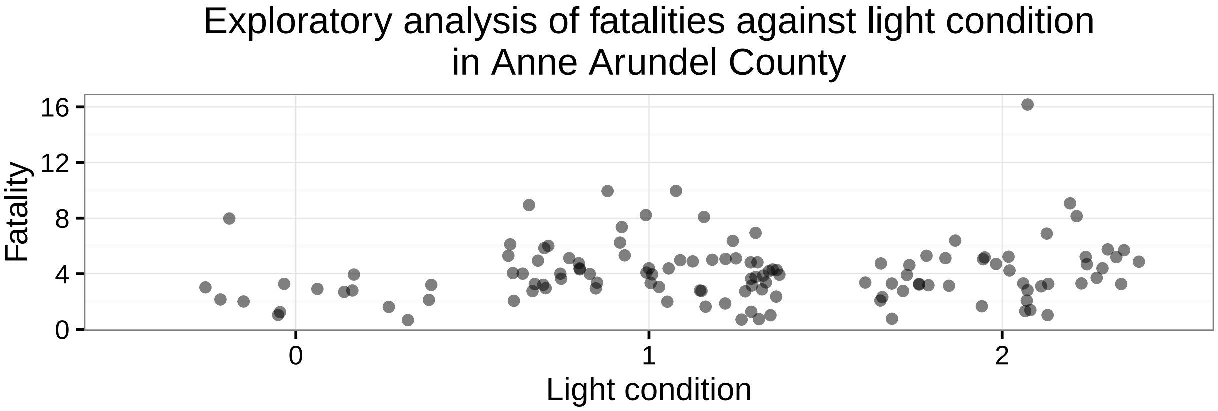


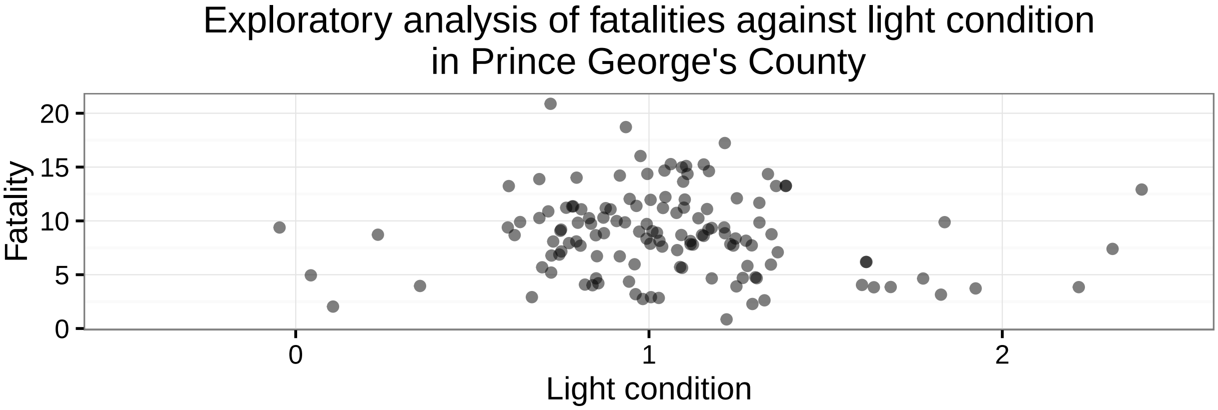
Thirdly, for our main focus, the unemployment rate for both cities reaches its peak around 2011 when the United States had entered a severe economic recession since 2008. The green line is a loess smoothing line drawn by default in ggplot and shaded area is confidence interval.


Exploratory analysis of trend of unemployment rate
in Anne Arundel County

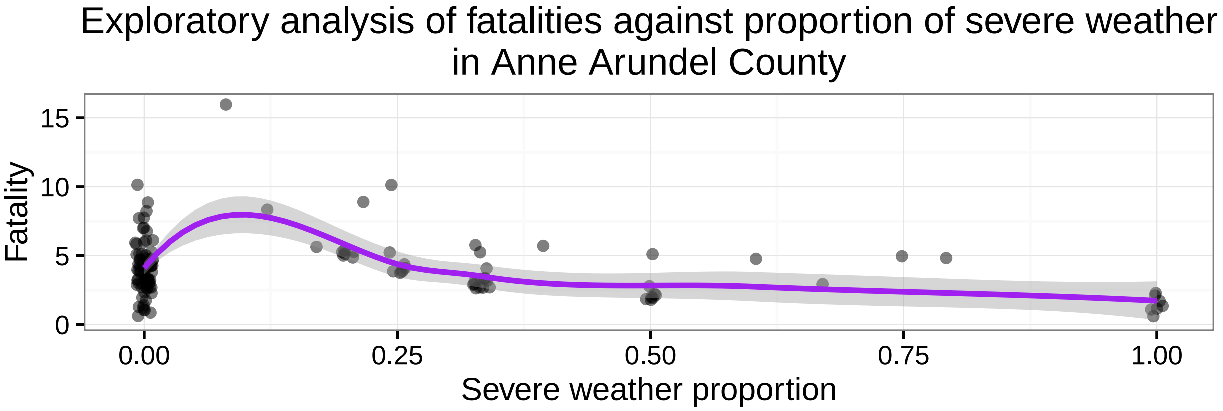

Exploratory analysis of trend of unemployment rate
in Prince George's County

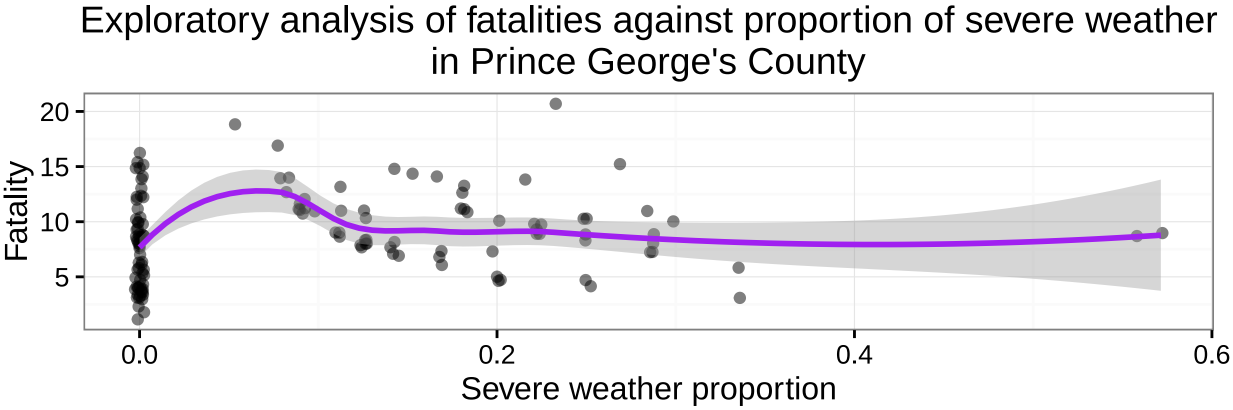
Furthermore, as we see from the exploratory plots of light conditions, we noticed that most of the fatal crashes happened when lights are obscure (i.e. during dawn, dusk or night with lights on).





Lastly, we plotted fatalities against weather condition. The purple lines are a loess smoothing line and the shaded area is point-wise confidence interval. We noticed that most fatal accidents happened during clear weather rather than severe weather.





Variable Selection

In this section, we will discuss how to include variables in details. lme4 package displays the Wald Z-statistics for each parameter in the model summary. They are convenient to compute but based on asymptotic approximations, assuming both that (1) the sampling distributions of the parameters are multivariate normal and that (2) the sampling distribution of the log-likelihood is (proportional to) χ^2 [1]. Based on some warning message from model, we decided to rescale the variables to obtain identifiable model.

To test effects, we used likelihood ratio test to see whether it's necessary to include terms such as random slope of unemployment within county with correlated intercept. Since p-value for random slope for unemployment rate is P=0.193, we did not include random slope for unemployment rate in our analysis.

In order to obtain unbiased estimates of coefficients and standard errors, we divided dataset into training set (50%) and test set (50%). As for testing parameters in GLMM, we chose to perform likelihood ratio test on each potential main effects as well as interaction terms and included them if they're statistically significant on training set. Then we obtained unbiased estimates of coefficients and its standard deviation on test set reported in our report.

Firstly, we observed that light conditions (P=4.23×10^{-15}),age (P=4.52×10^{-14}) and county-level population (P=8.44×10^{-9}) are statistically significant while weather (P=0.195) and drunk proportion (P=0.478) are not statistically significant. Here, since we notice seasonal pattern in exploratory analysis, we need to include a natural spline of calendar time with two degrees of freedom to account for seasonality based on our prior knowledge. Therefore, we included age, light conditions, calendar time and county-level population in our model.

Secondly, we testify significance of interaction terms. The interactions between unemployment rate and light condition (P=0.997), population and light condition (P=0.491), population and calendar time (P=0.21), light condition and weather (P=0.453) , light condition and age (P=0.363) and unemployment rate and age (P=0.482) are not statistically significant and therefore we did not include them in our model.

Model evaluation

In this section, we will check overdispersion for our model, calculate an analogue of a coefficient of determination (R^2) and include summary results for both fixed effects and radom effects.

Model

Let μ_i be the average traffic-related fatalities for i th county (i=1,2,$\cdots$, 24). Our final model is:

$$\log\mu_i = (\beta_0 + \gamma_i) + \beta_1 unemprate + \beta_2 lgtcond + \beta_3 population + \beta_4 ns(ym, 2) + \beta_5 age + \beta_6 (age - 40)^+$$

Here, β_0 is fixed intercept and $\beta_1 \cdots, \beta_6$ are fixed effects for unemployment rate, light condition, population, calendar time and age with its additional linear spline term respectively. γ_i represents additive random effect for county.

Fixed effects

The following table shows the estimated coefficients for fixed effects and their 95% confidence interval based on asymptotic approximation.

Table 2: Summary of coefficients' estimates and its 95% confidence intervals for fixed effects

	Estimates of coefficients	Standard Error	Lower CI	Upper CI
Intercept	1.0083	0.0898	0.8323	1.1842
unemployment rate	-0.0387	0.0150	-0.0682	-0.0092
light condition1	0.3520	0.0632	0.2281	0.4758
light condition2	0.0873	0.0650	-0.0400	0.2147
population	0.3662	0.0417	0.2846	0.4479
ns(year-month,2)1	-0.1586	0.1586	-0.4695	0.1523
ns(year-month,2)2	-0.1847	0.0636	-0.3094	-0.0600
age	0.3638	0.0536	0.2587	0.4689
$(age - 40)^+$	-0.4195	0.0554	-0.5280	-0.3109

In order to interpret results more clearly, we performed an exponential transformation on the estimated coefficients and the confidence intervals as shown in the table 3.

Table 3: Summary of exp(coef) and its 95% confidence intervals for fixed effects

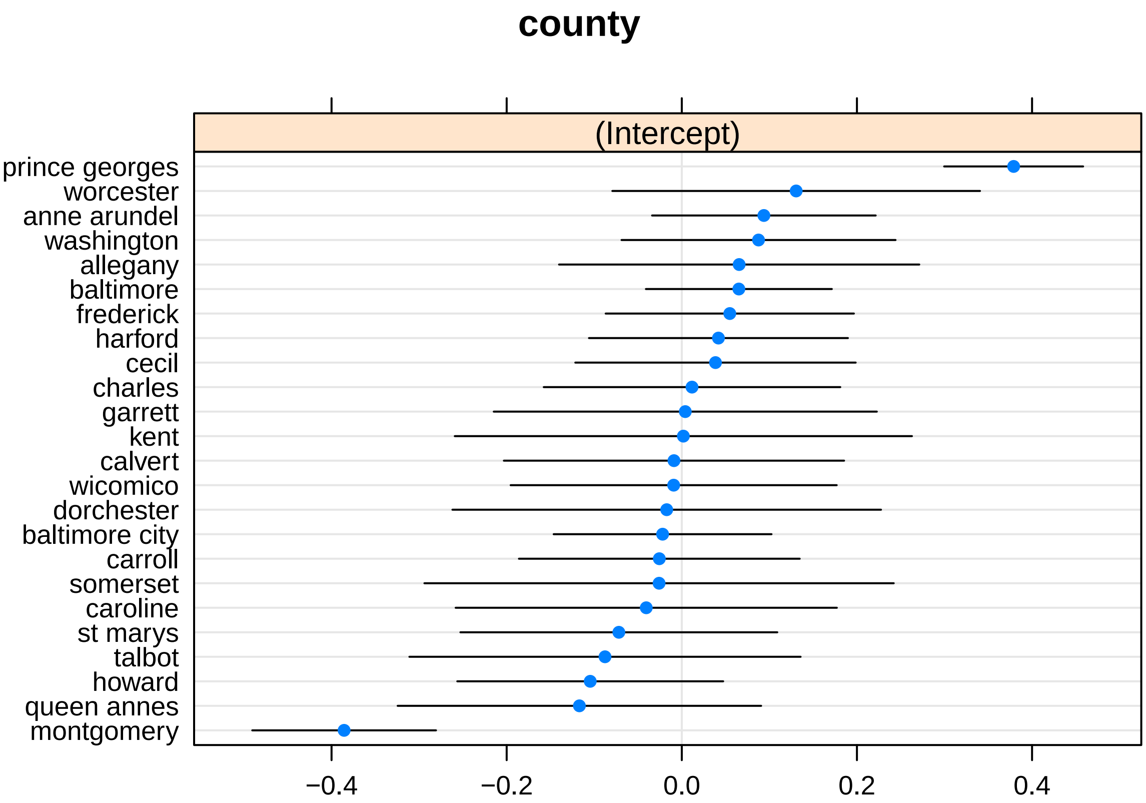
	Estimates of exp(coef)	Lower CI of exp(coef)	Upper CI of exp(coef)
Intercept	2.7408	2.2987	3.2680
unemployment rate	0.9620	0.9341	0.9908
light condition1	1.4219	1.2562	1.6094
light condition2	1.0913	0.9608	1.2394
population	1.4423	1.3292	1.5650
ns(year-month,2)1	0.8533	0.6253	1.1645
ns(year-month,2)2	0.8313	0.7339	0.9418
age	1.4388	1.2952	1.5982
$(age - 40)^+$	0.6574	0.5898	0.7328

At same level of light condition, population size, age and date (i.e. year-month), the expected traffic-related fatalities with 1 percent increase in unemployment rate is 0.962 times (95%CI: (0.9341, 0.9908)) than that with original unemployment rate. In other words, after controlling for age, population size, light condition and seasonality, the expected traffic-related fatalities decreases 3.8 % (95%CI: (0.92 %,6.59%)) with 1 percent increase in unemployment rate. For drivers who are younger than 40 years old, the expected counts of traffic fatalities increases 43.88% (95%CI: (29.52%, 59.82%)) with 1 unit increase in age. Similarly, for drivers that are older than 40 years old, 1 unit increases in age is associated with additional 43.88 % decreses (95%CI: (59.82 %,29.52%)) in expected traffic fatalities.

Random effects

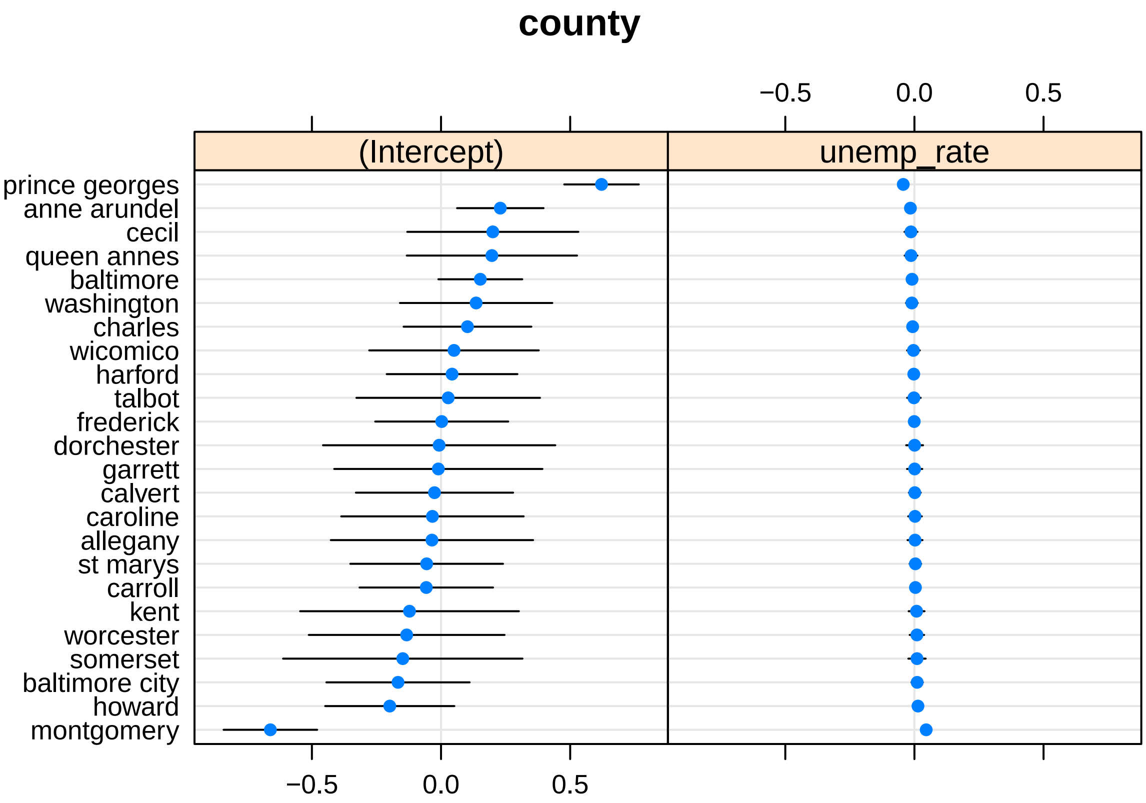
For random intercept at county level, we plotted estimated random intercepts and their confidence interval for all 24 counties in the above figure. Prince George's County has the largest radom county intercept while Montogomery County has the smallest value. This plot reveals difference across random intercept at county level and therefore proved necessity to include the random county effects.

```
## $county
```



If we also included random slope for unemployment rate, we plotted the estimated random effects and their confidence interval and discovered that there is no such big difference across counties for random slope of unemployment rate at county level. Thus, it is reasonable not to include random slope at county level.

```
## $county
```



Overdispersion

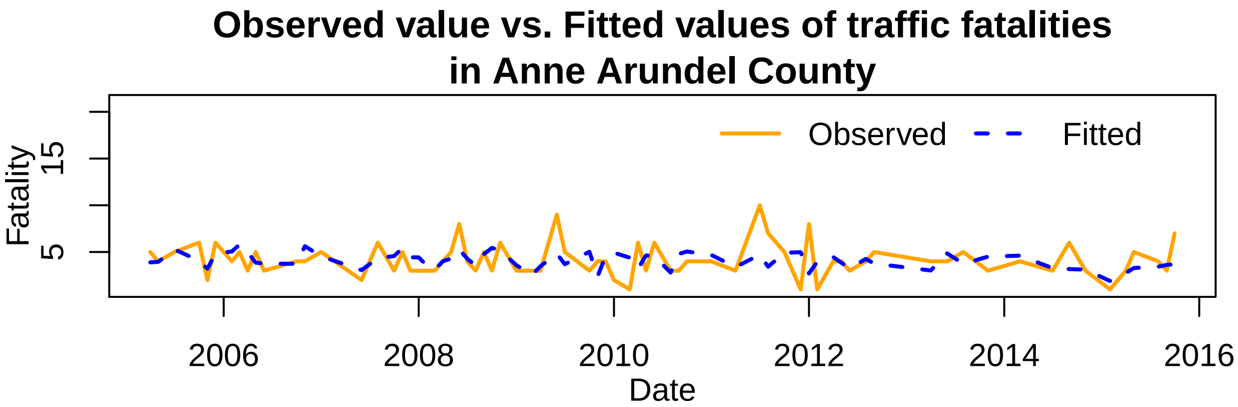
Using methodology and code provided in the FAQ list for the r-sig-mixed-models mailing list [1], our model did not have overdispersion problem. The method attempted to count every variance or covariance parameter as one degree of freedom and calculated p-value based on approximate chi-squared distribution. But it may not be an accurate result for overdispersion testing.

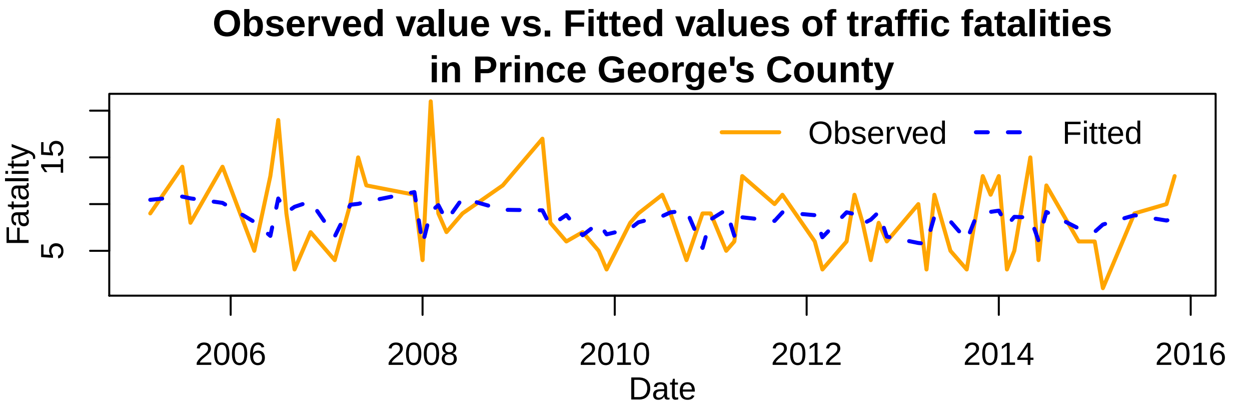
R-squared

Since it is another challenging question to get an analogue of R^2 or another simple goodness-of-fit metric for LMMs or GLMM, various quantification methods are proposed online [1]. Among those, we decided to use code provided by Jarrett Byrnes as one way to calculate coefficient of determination and R^2 for our model is 0.648.

Fitness of model

In report, we included results for baltimore city to illustrate fitness of our model. Here, we provided additional counties, namely Anne Arundel's County and Prince George's County, to illustrate the result. We fitted our model on test set (another 50%) and obtained fitted values. In the following figure, we plotted observed traffic fatalities and fitted values together to check fitness of our model. The orange solid line represents observed traffic fatalities in either counties and the blue dashed line represents the fitted traffic fatalities. As we could observe, our model captures the major seasonal trend and predict a majority of fatalities quite good but not with extreme cases.





All statistical analysis we mentioned above could reproduce in "final code.rmd" file.

Reference

[1] DRAFT r-sig-mixed-models FAQ. http://glmm.wikidot.com/faq

[2] FIPS county code - Wikipedia. https://en.wikipedia.org/wiki/FIPS_county_code

[3] Original S code by Richard A. Becker, Allan R. Wilks.R version by Ray Brownrigg. Enhancements by Thomas P Minka and Alex Deckmyn. (2016). maps: Draw Geographical Maps. R package version 3.1.1. https://CRAN.R-project.org/package=maps

[4] Fatality Analysis Reporting System (FARS) Analytical User's Manual 1975-2015