Predicting Collision Fatalities
Chris Pecaut – April 20, 2016

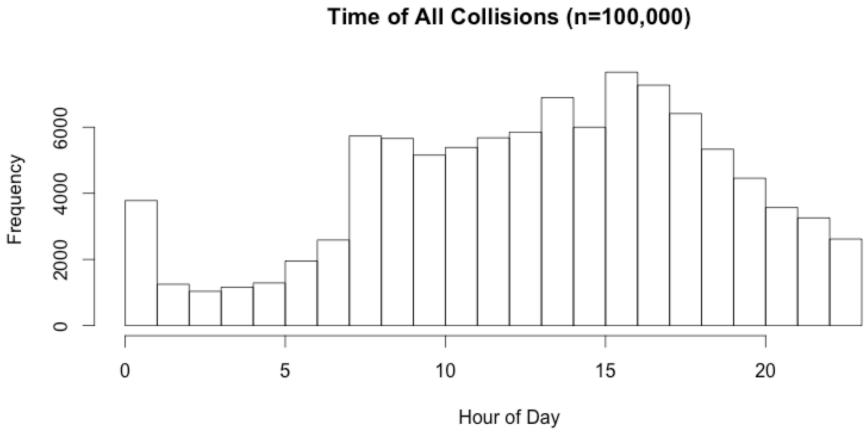
First I looked over the data to determine the year distribution. There are three complete sets of data from 2013, 2014, 2015 – with incomplete years in 2012 and 2016.

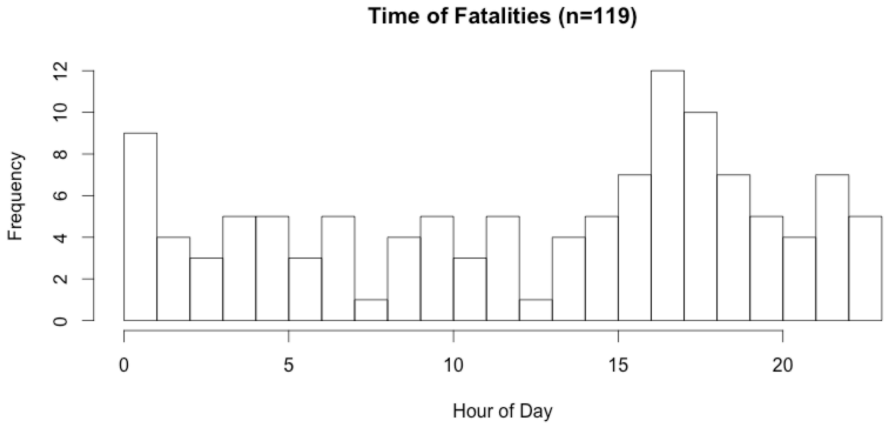
```
> table(fatal$year)

2012 2013 2014 2015
  18   39   34   28
> table(p2$year)

 2012  2013  2014  2015  2016
13696 27994 28325 29424   561
```

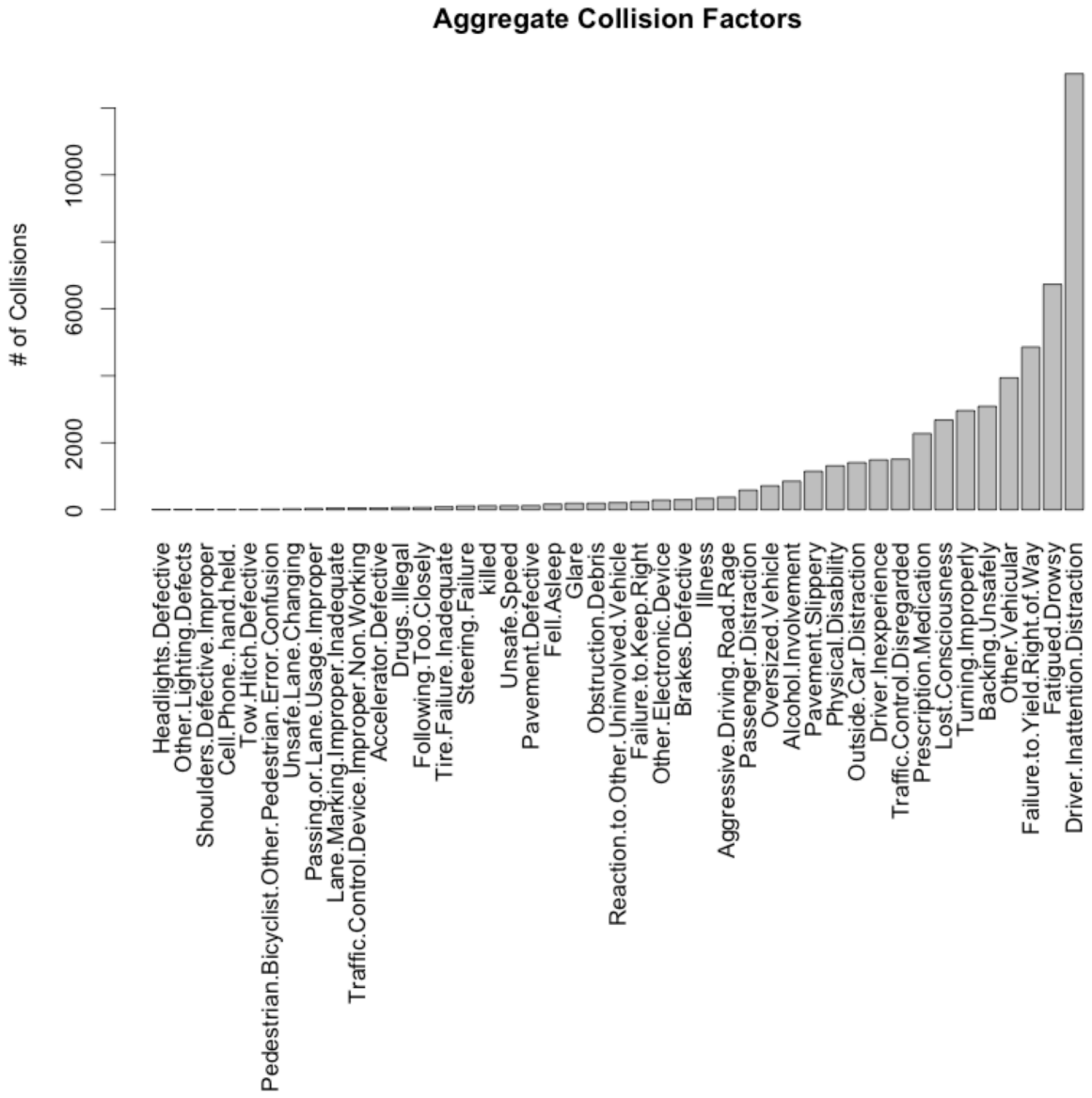
Since the data set had hour of day, I thought to break down the collisions by hour and compare the overall data set with the fatalities. The collisions and fatalities are normally distributed around the afternoon – with a spike in the early morning hour after midnight.





Question 2b) Looking at the data available, it seemed that building a model surrounding the collision factors was the most promising direction for predicting fatalities. As illustrated by the graph below of the prevalence of the 48 different collision factors, the frequency was evenly distributed across more than 15 of these factors.

The central challenge that I faced in this data set was the 5 different collision factor fields available for input, with many (> 12,000) having more than 2 factors, with the prevalence of duplicate factors across 1-5 columns.



So in preparation for running a logit regression to estimate the factors most closely associated with fatalities, I decided to convert the 5 colliison factor fields into dummy variables and then combine the 5 factors into a single, common field.

This process captured all of the unique collision factor information across all 5 fields, and eliminated any duplicates, creating a foundation for better accuracy to the logit model.

```
> summary(mylogit)

Call:
glm(formula = killed ~ ., family = binomial(link = "logit"),
    data = model)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5718  -0.0486  -0.0486  -0.0347   4.1980

Coefficients:
                                                  Estimate Std. Error z value             Pr(>|z|)
(Intercept)                                        -6.7428     0.1217 -55.394 < 0.0000000000000002 ***
Accelerator.Defective                             -14.8266  3960.2855  -0.004               0.9970
Aggressive.Driving.Road.Rage                        0.7580     1.0083   0.752               0.4522
Alcohol.Involvement                                 1.8136     0.4257   4.260   0.0000204523002495 ***
Backing.Unsafely                                    0.1286     0.5124   0.251               0.8018
Brakes.Defective                                  -14.7696  1679.6540  -0.009               0.9930
Cell.Phone..hand.held.                            -14.8289  9071.9248  -0.002               0.9987
Driver.Inattention.Distraction                     -0.6748     0.3701  -1.823               0.0682 .
Driver.Inexperience                                -0.5500     1.0059  -0.547               0.5845
Drugs..Illegal                                    -14.7322  3518.0282  -0.004               0.9967
Failure.to.Keep.Right                             -14.9788  1820.3609  -0.008               0.9934
Failure.to.Yield.Right.of.Way                      0.2819     0.3512   0.803               0.4221
Fatigued.Drowsy                                   -2.0686     1.0059  -2.056               0.0397 *
Fell.Asleep                                        1.0950     1.0553   1.038               0.2994
Following.Too.Closely                              3.6376     0.6161   5.904   0.0000000035482446 ***
Glare                                             -14.8340  2091.2204  -0.007               0.9943
Headlights.Defective                              -14.8233 10335.2276  -0.001               0.9989
Illness                                           -14.8137  1547.2456  -0.010               0.9924
Lane.Marking.Improper.Inadequate                  -14.6938  4204.0004  -0.003               0.9972
Lost.Consciousness                                -14.7677   554.4148  -0.027               0.9787
Obstruction.Debris                                -14.8589  2062.2410  -0.007               0.9943
Other.Electronic.Device                           -14.7849  1679.4891  -0.009               0.9930
Other.Lighting.Defects                            -14.7964 10304.2554  -0.001               0.9989
Other.Vehicular                                    -0.8172     0.7142  -1.144               0.2525
Outside.Car.Distraction                            -0.3576     1.0065  -0.355               0.7224
Oversized.Vehicle                                 -14.7051  1069.5677  -0.014               0.9890
Passenger.Distraction                              1.9914     0.4630   4.301   0.0000169685417235 ***
Passing.or.Lane.Usage.Improper                    -14.7447  4966.8564  -0.003               0.9976
Pavement.Defective                                -14.8241  2616.3747  -0.006               0.9955
Pavement.Slippery                                  -0.2830     1.0069  -0.281               0.7787
Pedestrian.Bicyclist.Other.Pedestrian.Error.Confusion -15.0707  7623.4216  -0.002               0.9984
Physical.Disability                                0.9819     0.4622   2.125               0.0336 *
Prescription.Medication                            -1.0508     1.0059  -1.045               0.2962
Reaction.to.Other.Uninvolved.Vehicle              -14.7203  1946.7340  -0.008               0.9940
Shoulders.Defective.Improper                      -14.6496 10129.7159  -0.001               0.9988
Steering.Failure                                  -14.7458  2786.9181  -0.005               0.9958
Tire.Failure.Inadequate                           -14.7300  3093.2845  -0.005               0.9962
Tow.Hitch.Defective                               -14.7383  8324.9098  -0.002               0.9986
Traffic.Control.Device.Improper.Non.Working       -14.8205  4002.2267  -0.004               0.9970
Traffic.Control.Disregarded                        2.1726     0.2812   7.725   0.0000000000000112 ***
Turning.Improperly                                -14.8192   525.4369  -0.028               0.9775
Unsafe.Lane.Changing                              -14.8187  6024.5047  -0.002               0.9980
Unsafe.Speed                                      -14.7044  2619.0619  -0.006               0.9955
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

I converted the results of the logit regression into a logit scalar so that the results can be interpreted straightforwardly. We can state the following from the results of the model:

For collisions classifed as:

> “Following Too Closely”, that increases the likelihood of a fatality by 0.43%.

"Traffic Control Disregarded", that increases the likelihood of fatality by 0.26%
"Passenger Distraction", that increased the likelihood of fatality by 0.24%
"Alcohol Involvement", that increases the likelihood of fatality by 0.21%
"Fatigued/Drowsy", that increases the likelihood of fatality by 0.13%
"Physical Disability", that increases the likelihood of fatality by 0.12%

Collision Factor	Total # of Collisions	Total # of Fatalities
Following Too Closely	67	3
Traffic Control Disregarded	1508	15
Passenger Distraction	584	5
Alcohol Involvement	854	6
Fatigued/Drowsy	6732	1
Physical Disability	1311	5
Failure to Yield	4860	9
Driver Inattention	13022	8
Backing Unsafely	3089	4

I chose to use a logit model because the explanatory variable in this case (collision fatality) was dichotomous, and the regression would provide a percentage value for the increase in risk with associated collision factors.

The results are basically intuitive in my judgment. Alcohol, Fatigue (falling asleep), and Passenger Distraction are reasonably associated with serious collisions.

Two of the collision factors with the highest number of fatalities, that did not result as statistically significant regressors, Driver Inattention, and Backing Unsafely (as depicted in table above) would reasonably result in more minor collisions.

Traffic Control Disregarded would likely involve high-speed collisions, and thereby more fatalities. By comparison with Failure to Yield, which would perhaps have involved lower speed collisions.

Question 2c) I would request that the severity of the collision be included, if possible in the data. Many accidents, despite being very severe, do not result in fatalities. And some of the categories with fewer collisions, such as the Headlights Defective, may rarely result in serious collisions.

The severity of the collision, by whatever cause, may be more decisive than the factor that caused the collision. And that severity may be more closely correlated with fatalities.

Another useful piece of data would be injuries, with a scale of the severity of injuries. Depending on the nature of the injury, whether a collision turned into a fatality could have hinged upon the time for an ambulance response, or the personal characteristics of the victim regarding age or health. I would request additional data on the victims and integrate that with the collision factors.