Airbag and other influences on accident fatalities

Toshov Nodirjon

California State University East Bay

1. Introduction

Car accident is one of the leading causes of death in the US. It is important to have some data analysis to understand the issue. The goal of this analysis is to answer the questions: What are the main factors that influence accident fatality? How do speed of impact, airbag, seatbelt, weight, age and other accident related factors affect accident fatality?

Two primary groups may benefit from this study. The first group, the consisting of drivers in US, may learn to identify factors that affect accident fatalities. By sharing this knowledge, drivers know how to reduce accident fatalities. Finally, educators can use these findings as a valuable guide to incorporate into their curriculum. By informing students the importance of developing programs to deal with accidents, the students may be able to transfer this knowledge to real life, thereby improving the quality of their attention while driving.

2. Data Description

Project data is the nassCDS data from DAAG package in R. Data is from police-reported car crashes in which there is a harmful event (people or property), and from which at least one vehicle was towed. Data is restricted to front-seat occupants, include only a subset of the variables recorded, and are restricted in other ways (DAGG). Originally, the data has 26217 observations and 15 variables. Response variable is dead, and potential predictor variables can be dvcat **(**speed of impact), weight, airbag, seatbelt, frontal, sex, age, age of occupant in years (ageOFocc), abcat(airbag availability and deploy status), and occRole. After skimming the data, 153 rows with missing data are removed, making project data to have 26063 rows. A quick summary of data set is shown in table 1 and table 2

Table 1

```
Skim summary statistics
 n obs: 26063
 n variables: 16

-- Variable type:character -------------------------------------------------------
 variable missing complete     n min max empty n_unique
   caseid       0    26063 26063   5   8     0     9400

-- Variable type:factor ----------------------------------------------------------
 variable missing complete     n n_unique                                  top_counts ordered
    abcat       0    26063 26063        3     una: 11727, dep: 8799, nod: 5537, NA: 0   FALSE
   airbag       0    26063 26063        2               air: 14336, non: 11727, NA: 0   FALSE
     dead       0    26063 26063        2                ali: 24883, dea: 1180, NA: 0   FALSE
    deadF       0    26063 26063        2                    0: 24883, 1: 1180, NA: 0   FALSE
   deploy       0    26063 26063        2                    0: 17264, 1: 8799, NA: 0   FALSE
    dvcat       0    26063 26063        5 10-: 12766, 25-: 8165, 40-: 2965, 55+: 1491    TRUE
  frontal       0    26063 26063        2                    1: 16775, 0: 9288, NA: 0   FALSE
  occRole       0    26063 26063        2                dri: 20541, pas: 5522, NA: 0   FALSE
 seatbelt       0    26063 26063        2                bel: 18465, non: 7598, NA: 0   FALSE
      sex       0    26063 26063        2                   m: 13885, f: 12178, NA: 0   FALSE

-- Variable type:numeric ---------------------------------------------------------
    variable missing complete     n    mean      sd   p0     p25     p50     p75     p100     hist
    ageOFocc       0    26063 26063   37.22   17.9    16   22      33      48       97     ▆▇▅▃▂▁
 injSeverity       0    26063 26063    1.72    1.29    0    1       2       3        6     ▇▃▂▁▁▁
      weight       0    26063 26063  462.48 1527.78    0   32.38   86.99  363.35 57871.59  ▇▁▁▁▁▁
     yearacc       0    26063 26063 1999.55    1.7  1997 1998    2000    2001     2002     ▇▁▇▁▇
     yearVeh       0    26063 26063 1992.8     5.59 1953 1989    1994    1997     2003     ▁▁▁▁▃▇
```

Table 2

	dvcat <ord>	weight <dbl>	dead <fctr>	airbag <fctr>	seatbelt <fctr>	frontal <fctr>	sex <fctr>	ageOFocc <dbl>	yearacc <dbl>
1	25-39	25.069	alive	none	belted	1	f	26	1997
2	10-24	25.069	alive	airbag	belted	1	f	72	1997
3	10-24	32.379	alive	none	none	1	f	69	1997
4	25-39	495.444	alive	airbag	belted	1	f	53	1997
5	25-39	25.069	alive	none	belted	1	f	32	1997
6	40-54	25.069	alive	none	belted	1	f	22	1997

frontal <fctr>	sex <fctr>	ageOFocc <dbl>	yearacc <dbl>	yearVeh <dbl>	abcat <fctr>	occRole <fctr>	deploy <fctr>	injSeverity <dbl>	caseid <chr>
1	f	26	1997	1990	unavail	driver	0	3	2:3:1
1	f	72	1997	1995	deploy	driver	1	1	2:3:2
1	f	69	1997	1988	unavail	driver	0	4	2:5:1
1	f	53	1997	1995	deploy	driver	1	1	2:10:1
1	f	32	1997	1988	unavail	driver	0	3	2:11:1
1	f	22	1997	1985	unavail	driver	0	3	2:11:2

The response variable of the analysis is “dead”. This is a factor variable with unbalance ratio of alive and dead (shown in figure 1)

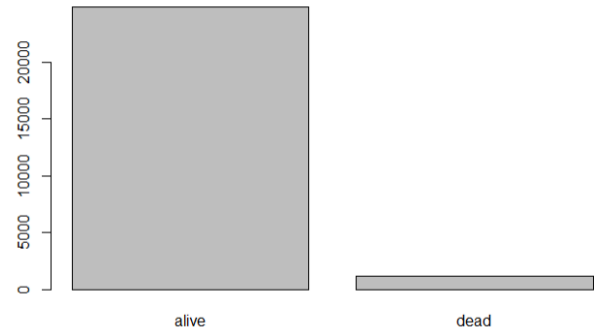


Figure 1

Other predictor variables could be dvcat, weight, airbag, seatbelt, frontal, sex, age, ageOFocc, abcat and occRole. Some variable are factor type and others are numeric type as shown in table 1 and table 2

3. Methods and Results

This analysis will mainly use logistic regression model, backward stepwise election, AIC for model selection and cross validation (Fox, 2019).

First, data is divided into two groups: the training group with 18244 rows, accounting for 70% of data and the test group with 7819, accounting for 30% of data (Table 3). Rows are selected randomly by R

Table 3

Group	Count	%
Training Group	18244	70%
Test Group	7819	30%
Total	26063	

Next, the first logistic regression model is computed by glm function on the training data set. This model has 10 predictor variables as shown below:

glm1 <- glm(dead ~ dvcat+ weight+ airbag+ seatbelt+ ageOFocc+ sex+ frontal+ yearacc+ abcat+ occRole+yearVeh, data=Airbag, subset=train, family=binomial)

Table 4

```
Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)    -8.034e+01  4.877e+01  -1.647   0.0995 .
dvcat.L         3.165e+00  3.780e-01   8.373   <2e-16 ***
dvcat.Q         6.486e-01  3.196e-01   2.029   0.0424 *
dvcat.C        -4.426e-01  2.059e-01  -2.149   0.0316 *
dvcat^4         1.141e-01  1.104e-01   1.033   0.3016
weight         -4.027e-03  4.549e-04  -8.851   <2e-16 ***
airbagairbag   -2.136e-01  1.262e-01  -1.693   0.0904 .
seatbeltbelted -9.087e-01  8.272e-02 -10.986   <2e-16 ***
ageOFocc        3.194e-02  2.101e-03  15.204   <2e-16 ***
sexm            1.533e-01  8.387e-02   1.828   0.0675 .
frontal1       -1.114e+00  8.752e-02 -12.730   <2e-16 ***
yearacc         2.616e-02  2.430e-02   1.077   0.2817
abcatnodeploy  -1.847e-01  1.387e-01  -1.331   0.1830
abcatunavail           NA         NA      NA       NA
occRolepass     1.821e-01  9.522e-02   1.913   0.0558 .
yearVeh         1.266e-02  1.046e-02   1.211   0.2260
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6733.1  on 18243  degrees of freedom
Residual deviance: 4690.3  on 18229  degrees of freedom
AIC: 4720.3
```

In the first model, dvcat, weight, seatbelt, ageOFocc, frontal and occRolepass are significant predictors of fatalities. On the other hand, airbag, sex, yearacc, and abcat are not significant predictors. Thus, backwards stepwise selection by step() function is used to conduct the second model with R code

```
glm2 <- step(glm1)
```

Result of coefficients are shown in table 5

Table 5

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -32.864185  20.509588  -1.602   0.1091
dvcat.L         3.166605   0.377984   8.378   <2e-16 ***
dvcat.Q         0.650043   0.319591   2.034   0.0420 *
dvcat.C        -0.442344   0.205908  -2.148   0.0317 *
dvcat^4         0.114274   0.110433   1.035   0.3008
weight         -0.004000   0.000453  -8.831   <2e-16 ***
seatbeltbelted -0.909390   0.082706 -10.995   <2e-16 ***
ageOFocc        0.031918   0.002101  15.195   <2e-16 ***
sexm            0.154894   0.083845   1.847   0.0647 .
frontal1       -1.109392   0.087392 -12.694   <2e-16 ***
abcatnodeploy  -0.179791   0.138636  -1.297   0.1947
abcatunavail    0.214863   0.126110   1.704   0.0884 .
occRolepass     0.180960   0.095193   1.901   0.0573 .
yearVeh         0.014978   0.010273   1.458   0.1448
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6733.1  on 18243  degrees of freedom
Residual deviance: 4691.4  on 18230  degrees of freedom
AIC: 4719.4
```

After backward stepwise selection, airbag, deploy and yearacc are removed. However, sex, abcat, and yearVeh are still insignificant predictors. Thus, the third model is computed by removing insignificant variables: abcat, sex, and yearVeh

glm3 <- glm(dead ~ dvcat + weight + seatbelt + ageOFocc + frontal + occRole, data=Airbag, subset=train, family=binomial).

Summary of coefficients are shown in table 7.Now, the best model is selected by comparing AIC (Table 3). Model with lowest AIC is the second model

Table 6

	df <dbl>	AIC <dbl>
glm1	15	4720.256
glm2	14	4719.415
glm3	10	4721.631

Table 7

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.8673747  0.1656493 -17.310   <2e-16 ***
dvcat.L         3.2272698  0.3766378   8.569   <2e-16 ***
dvcat.Q         0.6324042  0.3191136   1.982   0.0475 *
dvcat.C        -0.4399994  0.2058048  -2.138   0.0325 *
dvcat^4         0.1169148  0.1103722   1.059   0.2895
weight         -0.0040025  0.0004522  -8.852   <2e-16 ***
seatbeltbelted -0.9320438  0.0810643 -11.498   <2e-16 ***
ageOFocc        0.0314226  0.0020872  15.055   <2e-16 ***
frontal1       -1.0457870  0.0820816 -12.741   <2e-16 ***
occRolepass     0.1803324  0.0933971   1.931   0.0535 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6733.1  on 18243  degrees of freedom
Residual deviance: 4701.6  on 18234  degrees of freedom
AIC: 4721.6
```

Model 2 (Table 7) is now used in test data to predict probability of accident fatalities.

Logistic Regression equation in logit form:

logit(p(x)) = log (p(x) /(1 − p(x))) = β0 + β1dvcat + + β2weight + β3seatbelt + β4ageOFocc + β5sex + β6frontal + β7abcat + β8occRolepass + β9yearVeh

Coefficients of predictors are shown in Table 5

Confusion matrix, accuracy, sensitivity, and specificity are calculated as below

```
          actual
prediction alive dead  Sum
     alive  7437  308 7745
     dead     29   45   74
     Sum    7466  353 7819
```

```
# Accuracy (Percent Correctly Classified)
 (7437+45)/7819

## [1] 0.9568999

# Sensitivity (Percent dead Correctly Classified)
 45/353

## [1] 0.1274788

# Specificity (Precent alive Correctly Classified)
 7437/7466

## [1] 0.9961157
```

4. Discussion.

Probability of fatality is significantly influenced by impact speed, seatbelt, frontal impact, weight and age. In most cases, probability of accident fatality increases with higher impacted speed or older age. Probability of accident fatality decreases with seatbelt, higher weight, or frontal impact. The final logistic model has high accuracy of 95.6%, and high specificity of 99.6% while the low sensitivity 12.7% due to unbalanced data (proportion of dead over alive is 0.047)

Limitation of this analysis could be multicollinearity between predictor variables. For example, variable airbag, abcat and deploy are closely related regarding the airbag status. Diagnostic and assumption of logistic regression were not considered since they are out of scope of study. Future improvement of this study should focus on checking assumption of logistic regression and multicollinearity.

References

DAGG. *nassCDS*: *Airbag and other influences on accident fatalities*. (n.d.). Retrieved from https://rdrr.io/cran/DAAG/man/nassCDS.html

Fox, E. (2019). *Lecture 21: Logistic Regression [PDF file]*. Retrieved from CalState Eastbay Blackboard: http://www.csueastbay.edu

Appendix

ProjectAirbag r3

```
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```
library(skimr)
```

```
## Registered S3 method overwritten by 'skimr':
##   method      from
##   print.spark pillar
```

```
##
## Attaching package: 'skimr'
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
library(DAAG)
```

```
## Loading required package: lattice
```

```
library(tidyverse)
```

```
## Registered S3 method overwritten by 'rvest':
##   method            from
##   read_xml.response xml2
```

```
## -- Attaching packages ---------------------------------------------------
---------------------------------- tidyverse 1.2.1 --
```

```
## v tibble  2.1.1       v purrr   0.3.2
## v tidyr   0.8.3       v dplyr   0.8.0.1
## v readr   1.3.1       v stringr 1.4.0
## v tibble  2.1.1       v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------
---------------------------- tidyverse_conflicts() --
## x purrr::accumulate() masks foreach::accumulate()
## x tidyr::expand()     masks Matrix::expand()
## x dplyr::filter()     masks skimr::filter(), stats::filter()
## x dplyr::lag()        masks stats::lag()
## x purrr::when()       masks foreach::when()
```

```
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:DAAG':
##
##     cities
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
?nassCDS
```

```
## starting httpd help server ... done
```

```
head(nassCDS)
```

```
##   dvcat  weight  dead airbag seatbelt frontal sex ageOFocc yearacc yearVeh
## 1 25-39  25.069 alive   none   belted       1   f       26    1997    1990
## 2 10-24  25.069 alive airbag   belted       1   f       72    1997    1995
## 3 10-24  32.379 alive   none     none       1   f       69    1997    1988
## 4 25-39 495.444 alive airbag   belted       1   f       53    1997    1995
## 5 25-39  25.069 alive   none   belted       1   f       32    1997    1988
## 6 40-54  25.069 alive   none   belted       1   f       22    1997    1985
##     abcat occRole deploy injSeverity caseid
## 1 unavail  driver      0           3  2:3:1
## 2  deploy  driver      1           1  2:3:2
## 3 unavail  driver      0           4  2:5:1
## 4  deploy  driver      1           1 2:10:1
## 5 unavail  driver      0           3 2:11:1
## 6 unavail  driver      0           3 2:11:2
```

```
skim(nassCDS)
```

```
## Skim summary statistics
##  n obs: 26217
##  n variables: 15
##
## -- Variable type:character ----------------------------------------
-------------------------------------------------
##  variable missing complete     n min max empty n_unique
##     abcat       0    26217 26217   6   8     0        3
```

```
##    caseid       0    26217 26217   5   8     0     9409
##   occRole       0    26217 26217   4   6     0        2
##
## -- Variable type:factor ------------------------------------------------
-------------------------------------------------------
##  variable missing complete     n n_unique
##    airbag       0    26217 26217        2
##      dead       0    26217 26217        2
##     dvcat       0    26217 26217        5
##  seatbelt       0    26217 26217        2
##       sex       0    26217 26217        2
##                                  top_counts ordered
##                 air: 14419, non: 11798, NA: 0   FALSE
##                  ali: 25037, dea: 1180, NA: 0   FALSE
##  10-: 12848, 25-: 8214, 40-: 2977, 55+: 1492    TRUE
##                  bel: 18573, non: 7644, NA: 0   FALSE
##                     m: 13969, f: 12248, NA: 0   FALSE
##
## -- Variable type:numeric -----------------------------------------------
-------------------------------------------------------
##     variable missing complete     n    mean      sd   p0     p25     p50
##     ageOFocc       0    26217 26217   37.21   17.91   16   22      33
##       deploy       0    26217 26217    0.34    0.47    0    0       0
##      frontal       0    26217 26217    0.64    0.48    0    0       1
##  injSeverity     153    26064 26217    1.72    1.29    0    1       2
##       weight       0    26217 26217  462.81 1524.84    0   32.47   86.99
##      yearacc       0    26217 26217 1999.56    1.7  1997 1998    2000
##      yearVeh       1    26216 26217 1992.8     5.59 1953 1989    1994
##      p75     p100     hist
##    48       97    <U+2587><U+2585><U+2583><U+2582><U+2582><U+2581><U+2581>
<U+2581>
##     1        1    <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
<U+2585>
##     1        1    <U+2585><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
<U+2587>
##     3        6    <U+2586><U+2585><U+2583><U+2587><U+2581><U+2581><U+2581>
<U+2581>
##   364.72 57871.59 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
<U+2581>
##  2001     2002    <U+2587><U+2587><U+2581><U+2587><U+2587><U+2581><U+2587>
<U+2587>
##  1997     2003    <U+2581><U+2581><U+2581><U+2581><U+2581><U+2585><U+2587>
<U+2586>
```

Remove missing data and factor data

```
Airbag <-nassCDS
Airbag <- na.omit(Airbag)
Airbag$deploy <- as.factor(Airbag$deploy)
```

```
Airbag$frontal <- as.factor(Airbag$frontal)
Airbag$abcat <- as.factor(Airbag$abcat)
Airbag$dvcat <- as.factor(Airbag$dvcat)
Airbag$occRole <- as.factor(Airbag$occRole)

skim(Airbag)

## Skim summary statistics
##  n obs: 26063
##  n variables: 15
##
## -- Variable type:character -----------------------------------------------
-------------------------------------------------------
##  variable missing complete     n min max empty n_unique
##    caseid       0    26063 26063   5   8     0     9400
##
## -- Variable type:factor --------------------------------------------------
-------------------------------------------------------
##  variable missing complete     n n_unique
##     abcat       0    26063 26063        3
##    airbag       0    26063 26063        2
##      dead       0    26063 26063        2
##    deploy       0    26063 26063        2
##     dvcat       0    26063 26063        5
##   frontal       0    26063 26063        2
##   occRole       0    26063 26063        2
##  seatbelt       0    26063 26063        2
##       sex       0    26063 26063        2
##                                  top_counts ordered
##      una: 11727, dep: 8799, nod: 5537, NA: 0   FALSE
##                air: 14336, non: 11727, NA: 0   FALSE
##                 ali: 24883, dea: 1180, NA: 0   FALSE
##                     0: 17264, 1: 8799, NA: 0   FALSE
##  10-: 12766, 25-: 8165, 40-: 2965, 55+: 1491    TRUE
##                    1: 16775, 0: 9288, NA: 0   FALSE
##                 dri: 20541, pas: 5522, NA: 0   FALSE
##                 bel: 18465, non: 7598, NA: 0   FALSE
##                    m: 13885, f: 12178, NA: 0   FALSE
##
## -- Variable type:numeric -------------------------------------------------
-------------------------------------------------------
##     variable missing complete     n    mean      sd   p0     p25     p50
##     ageOFocc       0    26063 26063   37.22   17.9    16   22      33
##  injSeverity       0    26063 26063    1.72    1.29    0    1       2
##       weight       0    26063 26063  462.48 1527.78    0   32.38   86.99
##      yearacc       0    26063 26063 1999.55    1.7  1997 1998    2000
##      yearVeh       0    26063 26063 1992.8     5.59 1953 1989    1994
##      p75     p100     hist
##    48       97    <U+2587><U+2585><U+2583><U+2582><U+2582><U+2581><U+2581>
<U+2581>
```

```
##     3        6    <U+2586><U+2585><U+2583><U+2587><U+2581><U+2581><U+2581>
<U+2581>
##   363.35 57871.59 <U+2587><U+2581><U+2581><U+2581><U+2581><U+2581><U+2581>
<U+2581>
##  2001     2002    <U+2587><U+2587><U+2581><U+2587><U+2587><U+2581><U+2587>
<U+2587>
##  1997     2003    <U+2581><U+2581><U+2581><U+2581><U+2581><U+2585><U+2587>
<U+2586>
```

```
summary(Airbag)
```

```
##      dvcat           weight            dead          airbag
##  1-9km/h:  676   Min.   :    0.00   alive:24883   none  :11727
##  10-24  :12766   1st Qu.:   32.38   dead : 1180   airbag:14336
##  25-39  : 8165   Median :   86.99
##  40-54  : 2965   Mean   :  462.48
##  55+    : 1491   3rd Qu.:  363.35
##                  Max.   :57871.59
##    seatbelt     frontal   sex          ageOFocc        yearacc
##  none  : 7598   0: 9288   f:12178   Min.   :16.00   Min.   :1997
##  belted:18465   1:16775   m:13885   1st Qu.:22.00   1st Qu.:1998
##                                     Median :33.00   Median :2000
##                                     Mean   :37.22   Mean   :2000
##                                     3rd Qu.:48.00   3rd Qu.:2001
##                                     Max.   :97.00   Max.   :2002
##     yearVeh          abcat         occRole      deploy     injSeverity
##  Min.   :1953   deploy  : 8799   driver:20541   0:17264   Min.   :0.000
##  1st Qu.:1989   nodeploy: 5537   pass  : 5522   1: 8799   1st Qu.:1.000
##  Median :1994   unavail :11727                            Median :2.000
##  Mean   :1993                                             Mean   :1.716
##  3rd Qu.:1997                                             3rd Qu.:3.000
##  Max.   :2003                                             Max.   :6.000
##     caseid
##  Length:26063
##  Class :character
##  Mode  :character
##
##
##
```

```
head(Airbag)
```

```
##   dvcat  weight  dead airbag seatbelt frontal sex ageOFocc yearacc yearVeh
## 1 25-39  25.069 alive   none   belted       1   f       26    1997    1990
## 2 10-24  25.069 alive airbag   belted       1   f       72    1997    1995
## 3 10-24  32.379 alive   none     none       1   f       69    1997    1988
## 4 25-39 495.444 alive airbag   belted       1   f       53    1997    1995
## 5 25-39  25.069 alive   none   belted       1   f       32    1997    1988
## 6 40-54  25.069 alive   none   belted       1   f       22    1997    1985
##     abcat occRole deploy injSeverity caseid
## 1 unavail  driver      0           3  2:3:1
```

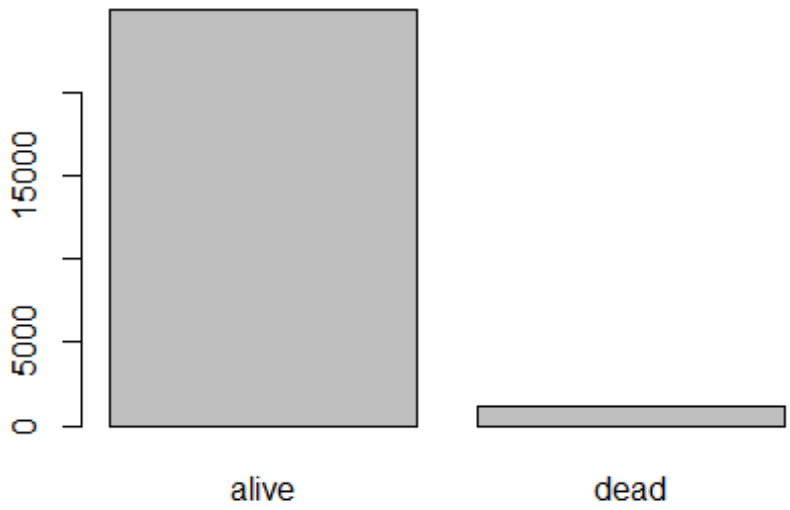
```
## 2  deploy  driver      1           1  2:3:2
## 3 unavail  driver      0           4  2:5:1
## 4  deploy  driver      1           1 2:10:1
## 5 unavail  driver      0           3 2:11:1
## 6 unavail  driver      0           3 2:11:2

contrasts(Airbag$dead)

##       dead
## alive    0
## dead     1
```

create extra variable for calculation

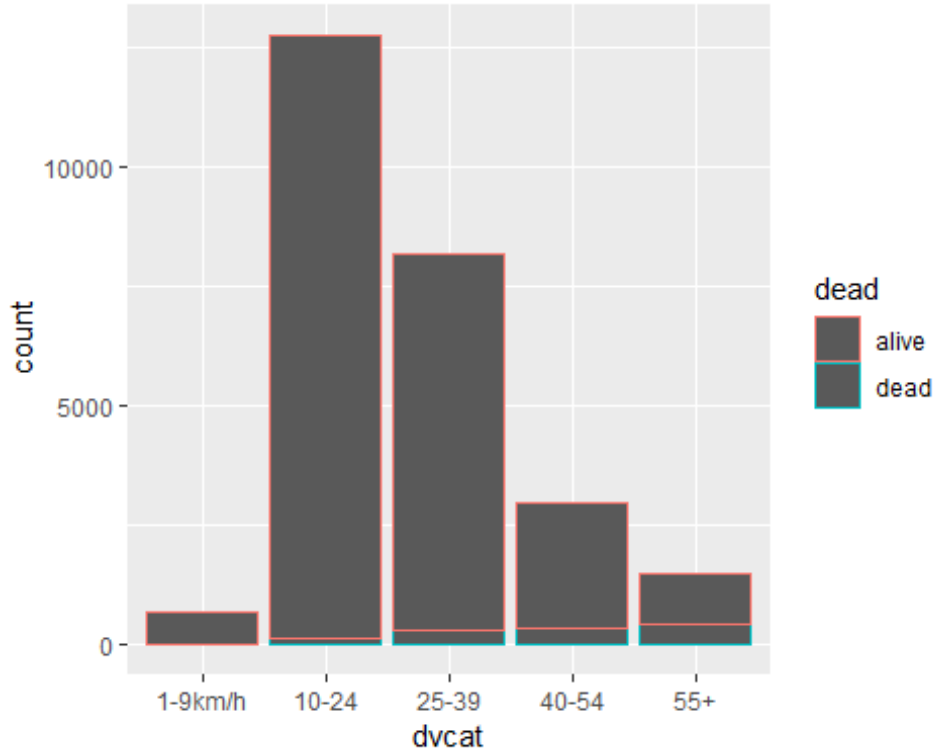
```
Airbag <-Airbag%>%mutate(deadF = if_else(dead=="alive",0,1))
Airbag$deadF <- as.factor(Airbag$deadF)
```

Descritive statistic for main variables

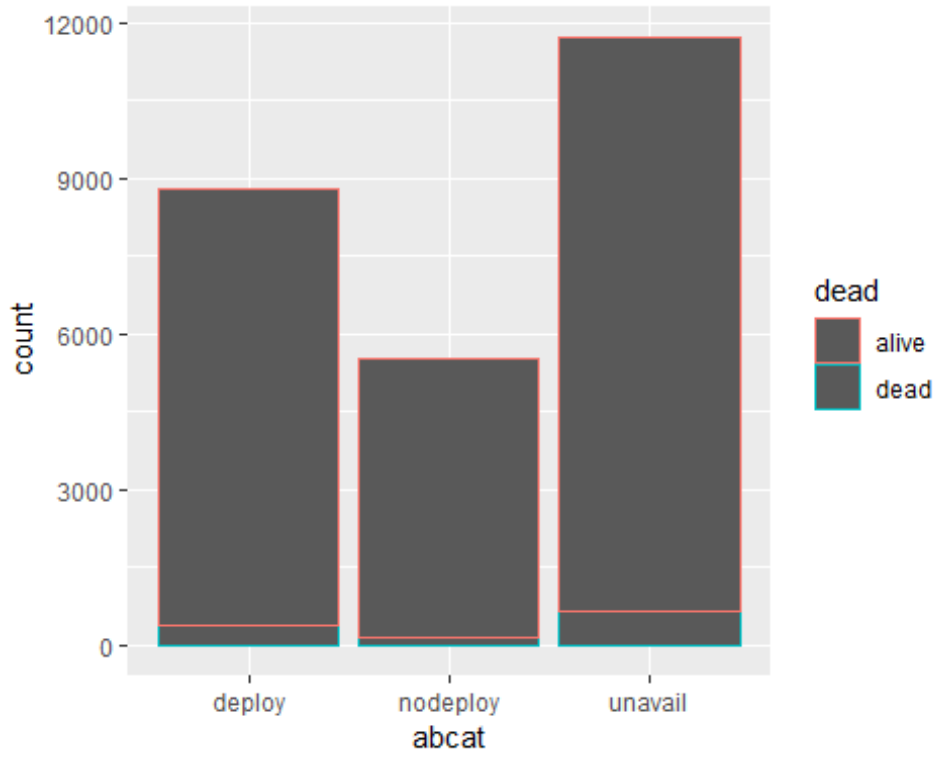
```
table(Airbag$dead)

##
## alive  dead
## 24883  1180

plot(Airbag$dead)
```



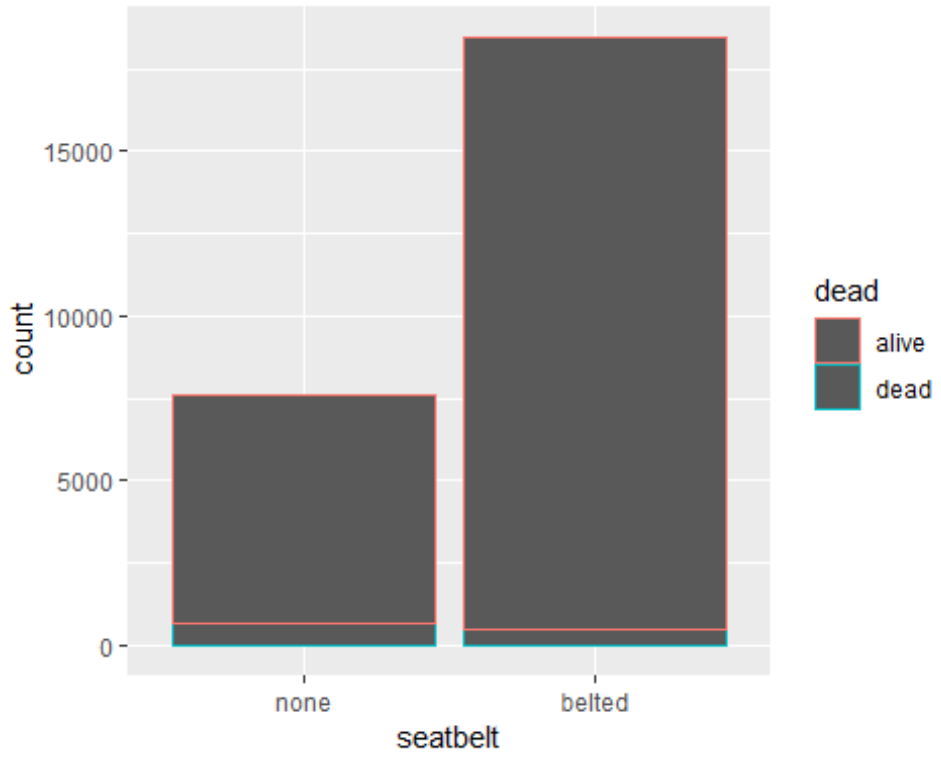
```r
#main predictor Variables
ggplot(Airbag, aes(dvcat, color=dead)) +  geom_bar() +
  scale_x_discrete(drop = FALSE)
```




```r
#main predictor Variables
ggplot(Airbag, aes(abcat, color=dead)) +
  geom_bar() +
  scale_x_discrete(drop = FALSE)
```




```
#main predictor Variables
ggplot(Airbag, aes(seatbelt, color=dead)) +
  geom_bar() +
  scale_x_discrete(drop = FALSE)
```



Divide data to train set and test set

```
n <- nrow(Airbag)
n

## [1] 26063

ntrain <-floor(0.7*n)
ntrain

## [1] 18244

set.seed(100)

floor(0.7*n)

## [1] 18244

train <- sample(1:n, ntrain)
```

Compute first model with the most predictor variables

```
glm1 <- glm(dead ~ dvcat+ weight+ airbag+ seatbelt+ ageOFocc+ sex+ frontal+ y
earacc+ abcat+ occRole+yearVeh,
data=Airbag, subset=train, family=binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(glm1)
```

```
##
## Call:
## glm(formula = dead ~ dvcat + weight + airbag + seatbelt + ageOFocc +
##     sex + frontal + yearacc + abcat + occRole + yearVeh, family = binomial
,
##     data = Airbag, subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8432  -0.2519  -0.1307  -0.0590   5.1985
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -8.034e+01  4.877e+01  -1.647   0.0995 .
## dvcat.L         3.165e+00  3.780e-01   8.373   <2e-16 ***
## dvcat.Q         6.486e-01  3.196e-01   2.029   0.0424 *
## dvcat.C        -4.426e-01  2.059e-01  -2.149   0.0316 *
## dvcat^4         1.141e-01  1.104e-01   1.033   0.3016
## weight         -4.027e-03  4.549e-04  -8.851   <2e-16 ***
## airbagairbag   -2.136e-01  1.262e-01  -1.693   0.0904 .
## seatbeltbelted -9.087e-01  8.272e-02 -10.986   <2e-16 ***
## ageOFocc        3.194e-02  2.101e-03  15.204   <2e-16 ***
## sexm            1.533e-01  8.387e-02   1.828   0.0675 .
## frontal1       -1.114e+00  8.752e-02 -12.730   <2e-16 ***
## yearacc         2.616e-02  2.430e-02   1.077   0.2817
## abcatnodeploy  -1.847e-01  1.387e-01  -1.331   0.1830
## abcatunavail           NA         NA      NA       NA
## occRolepass     1.821e-01  9.522e-02   1.913   0.0558 .
## yearVeh         1.266e-02  1.046e-02   1.211   0.2260
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6733.1  on 18243  degrees of freedom
## Residual deviance: 4690.3  on 18229  degrees of freedom
## AIC: 4720.3
##
## Number of Fisher Scoring iterations: 10
```

Backward stepwise selection by step function

```
glm2 <- step(glm1)
```

```
## Start:  AIC=4720.26
## dead ~ dvcat + weight + airbag + seatbelt + ageOFocc + sex +
##     frontal + yearacc + abcat + occRole + yearVeh
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=4720.26
## dead ~ dvcat + weight + seatbelt + ageOFocc + sex + frontal +
##     yearacc + abcat + occRole + yearVeh

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##            Df Deviance    AIC
## - yearacc   1   4691.4 4719.4
## - yearVeh   1   4691.7 4719.7
## <none>          4690.3 4720.3
## - sex       1   4693.6 4721.6
```

```
## - occRole   1   4693.9 4721.9
## - abcat     2   4696.9 4722.9
## - seatbelt  1   4812.2 4840.2
## - frontal   1   4850.8 4878.8
## - weight    1   4870.5 4898.5
## - ageOFocc  1   4917.9 4945.9
## - dvcat     4   5674.7 5696.7

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##
## Step:  AIC=4719.41
## dead ~ dvcat + weight + seatbelt + ageOFocc + sex + frontal +
##     abcat + occRole + yearVeh

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

##            Df Deviance    AIC
## <none>          4691.4 4719.4
## - yearVeh   1   4693.6 4719.6
## - sex       1   4694.8 4720.8
## - occRole   1   4695.0 4721.0
## - abcat     2   4698.0 4722.0
## - seatbelt  1   4813.5 4839.5
## - frontal   1   4851.0 4877.0
## - weight    1   4870.7 4896.7
## - ageOFocc  1   4918.8 4944.8
## - dvcat     4   5677.3 5697.3

summary(glm2)

##
## Call:
## glm(formula = dead ~ dvcat + weight + seatbelt + ageOFocc + sex +
##     frontal + abcat + occRole + yearVeh, family = binomial, data = Airbag,
##     subset = train)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8327  -0.2520  -0.1313  -0.0591   5.1949
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -32.864185  20.509588  -1.602   0.1091
## dvcat.L          3.166605   0.377984   8.378   <2e-16 ***
## dvcat.Q          0.650043   0.319591   2.034   0.0420 *
## dvcat.C         -0.442344   0.205908  -2.148   0.0317 *
## dvcat^4          0.114274   0.110433   1.035   0.3008
## weight          -0.004000   0.000453  -8.831   <2e-16 ***
## seatbeltbelted  -0.909390   0.082706 -10.995   <2e-16 ***
## ageOFocc         0.031918   0.002101  15.195   <2e-16 ***
## sexm             0.154894   0.083845   1.847   0.0647 .
## frontal1        -1.109392   0.087392 -12.694   <2e-16 ***
## abcatnodeploy   -0.179791   0.138636  -1.297   0.1947
## abcatunavail     0.214863   0.126110   1.704   0.0884 .
## occRolepass      0.180960   0.095193   1.901   0.0573 .
## yearVeh          0.014978   0.010273   1.458   0.1448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6733.1  on 18243  degrees of freedom
## Residual deviance: 4691.4  on 18230  degrees of freedom
## AIC: 4719.4
##
## Number of Fisher Scoring iterations: 10
```

P values of abcat and sex > 0.05. #Manually remove insignificant variables: abcat and sex

```
glm3 <- glm(dead ~ dvcat + weight + seatbelt + ageOFocc + frontal +  occRole,
data=Airbag, subset=train, family=binomial)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(glm3)

##
## Call:
## glm(formula = dead ~ dvcat + weight + seatbelt + ageOFocc + frontal +
##     occRole, family = binomial, data = Airbag, subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8552  -0.2512  -0.1321  -0.0602   5.2094
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)    -2.8673747  0.1656493 -17.310   <2e-16 ***
## dvcat.L         3.2272698  0.3766378   8.569   <2e-16 ***
## dvcat.Q         0.6324042  0.3191136   1.982   0.0475 *
## dvcat.C        -0.4399994  0.2058048  -2.138   0.0325 *
## dvcat^4         0.1169148  0.1103722   1.059   0.2895
## weight         -0.0040025  0.0004522  -8.852   <2e-16 ***
## seatbeltbelted -0.9320438  0.0810643 -11.498   <2e-16 ***
## ageOFocc        0.0314226  0.0020872  15.055   <2e-16 ***
## frontal1       -1.0457870  0.0820816 -12.741   <2e-16 ***
## occRolepass     0.1803324  0.0933971   1.931   0.0535 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 6733.1  on 18243  degrees of freedom
## Residual deviance: 4701.6  on 18234  degrees of freedom
## AIC: 4721.6
##
## Number of Fisher Scoring iterations: 10
```

Comparing AIC

```
AIC(glm1,glm2, glm3)
```

```
##      df      AIC
## glm1 15 4720.256
## glm2 14 4719.415
## glm3 10 4721.631
```

Model 2 has lowest AIC score. MOdel 2 is selected.

Cross Validation

```
 Airbag_test <- Airbag[-train, ]
 ntest <-nrow(Airbag_test)
 ntest
```

```
## [1] 7819
```

```
 probs_test <- predict(glm2, newdata = Airbag_test,type="response")
 preds_test <- rep("alive", ntest)
 preds_test[probs_test > 0.5] <- "dead"
 head(probs_test)
```

```
##            5            6            9           12           13
## 0.0142533623 0.0361854315 0.0007669697 0.0123997205 0.0003440137
##           17
## 0.0451277544
```

```
tb <- table(prediction = preds_test, actual = Airbag_test$dead)
addmargins(tb)
```

```
##           actual
## prediction alive dead  Sum
##      alive  7437  308 7745
##      dead     29   45   74
##      Sum    7466  353 7819
```

```
# Accuracy (Percent Correctly Classified)
 (7437+45)/7819
```

```
## [1] 0.9568999
```

```
# Sensitivity (Percent dead Correctly Classified)
 45/353
```

```
## [1] 0.1274788
```

```
# Specificity (Precent alive Correctly Classified)
 7437/7466
```

```
## [1] 0.9961157
```

```
Airbag%>%group_by(dead)%>% summarise(n=n(), pct = n/26063 )
```

```
## # A tibble: 2 x 3
##   dead      n    pct
##   <fct> <int>  <dbl>
## 1 alive 24883 0.955
## 2 dead   1180 0.0453
```

```
1180/24883
```

```
## [1] 0.04742193
```

```
citation("DAAG")
```

```
##
## To cite package 'DAAG' in publications use:
##
##   John H. Maindonald and W. John Braun (2019). DAAG: Data Analysis
##   and Graphics Data and Functions. R package version 1.22.1.
##   https://CRAN.R-project.org/package=DAAG
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {DAAG: Data Analysis and Graphics Data and Functions},
##     author = {John H. Maindonald and W. John Braun},
##     year = {2019},
##     note = {R package version 1.22.1},
##     url = {https://CRAN.R-project.org/package=DAAG},
```

```
##   }
##
## ATTENTION: This citation information has been auto-generated from
## the package DESCRIPTION file and may need manual editing, see
## 'help("citation")'.
```