Airbag and other influences on accident fatalities

The purpose of this project is to determine the impact of the Airbag and other predictors responsible for accident fatalities based on multiple methods.

STAT 694

Toshov Nodirjon

Research question

1. What are the main factors that influence accident fatality?

(How do airbag, speed of impact, airbag, seatbelt, weight, age and other factors affect accident fatality?)

2. How well does the model fit the data?

Data Description

- nassCDS data from DAAG package in R
  - US data, for 1997-2002,
  - from police-reported car crashes in which there is a harmful event
  - from which at least one vehicle was towed.
  - data are restricted to front-seat occupants
- 26217 observations, 15 variables
- Exclude 153 rows with missing data, making project data to have 26063 rows

Data Description

```
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
```

Data Description

Response Variable	Description
dead	factor with levels alive dead
Predictor Variables	Description
dvcat	ordered factor with levels (estimated impact speeds) 1-9km/h, 10-24, 25-39, 40-54,55+
abcat	Did one or more (driver or passenger) airbag(s) deploy? This factor has levels deploy, nodeploy, unavailable
ageOFocc	age of occupant in years
airbag	a factor with levels none airbag
dead	factor with levels alive dead
frontal	a numeric vector; 0 = non-frontal, 1=frontal impact
occRole	a factor with levels driver pass
seatbelt	a factor with levels none belted
sex	a factor with levels f m
weight	Observation weights, albeit of uncertain accuracy, designed to account for varying sampling probabilities.
yearacc	year of accident
yearVeh	Year of model of vehicle; a numeric vector

Method

Logistic Regression

- Response variable: **dead**
- Predictors variables: **dvcat, weight, airbag, seatbelt, frontal, sex, age, ageOFocc, abcat and occRole**

Backwards stepwise selection

AIC for model selection

Cross Validation

Method

Data is divided into training group and test group

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

Method

Compute first mode with the most number of predictor variables from training data

```
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
---
```

Method

Backwards stepwise selection by step() function

- glm2 <- step(glm1)
- summary(glm2)

```
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
---
```

Method

Manually remove insignificant variables: abcat, sex, and yearVeh

```
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 .
---
```

Method

Comparing AIC

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

Result

Probability of fatality is significantly influenced by:

- impact speed
- seatbelt
- location of injury
- weight
- Age of occupants

```
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
---
```

Result

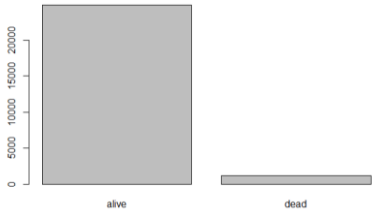
Cross validation by confusion matrix

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

```
# Accuracy (Percent Correctly Classified)
 (7437+339) / 7819

## [1] 0.9945006

# Sensitivity (Percent dead Correctly Classified)
 45/353

## [1] 0.1274788

# Specificity (Precent alive Correctly Classified)
 7437/7466

## [1] 0.9961157
```

Discussion

- High accuracy and specificity, but low sensitivity due to unbalanced data



dead <fctr>	n <int>	percent <dbl>
alive	24883	0.95472509
dead	1180	0.04527491

- Limitation of this analysis could be multicollinearity between predictor variables
- Diagnostic and assumption of logistic regression were not considered because it’s out of scope of study

Thank You!