Car Crash - Logistic Regression

Arthur L. Lui

Department of Statistics
Brigham Young University

March 27, 2014

Introduction

- More than 34,000 motor vehicle deaths in nation in 2012
- FHWA reponsible for improving roadway safety
- FHWA created FARS to collect relevant fatality data
- Goal: Understand relationship between independent variables and probability of fatality

Data

Response	Description
Fatal	1 - death in vehicle, 2 - no death in vehicle

Variable	Description
Year	Year of accident
DOW	Day of the week (1 =Sunday)
Hour	Hour at the time of accident
Mod	year Model year of vehicle involved in accident
Height	Driver height
Weight	Driver weight
DWI	Number of previous DWIs of driver
Age	Age of driver
Car.Type	Type of vehicle
Day	Day of the month
Drugs	Were drugs involved?
Drink	Had the driver been drinking?
Light	Light condition at time of accident
Month	Month of accident
Belt	Type of restraint used
Route	Type of highway
Sex	Gender of driver
Speed.Related	Was the accident speed related?
Speed.Limit	Posted speed limit
Road.Conditions	Condition of road at time of accident
Road.Type	Road type
Distracted	Was the driver distracted?

Data Cleaning

- Remove 99^{th} hour (23)
- Remove 9999 model year (4)
- Group model years < 1987 into 1986
- Remove Year variable (all are 2012)
- Change one No-Helmet death(1) to a survive(0)

Model

Can't use linear model beause:

- $Y_i \in \{0,1\} \Rightarrow \mathbf{Y}|\mathbf{X}$ not Normal
- $Y_i \in \{0,1\} \Rightarrow \mathbf{Y}|\mathbf{X} \sim \text{Bernoulli}(p_i)$

Logistic Regression Model

$$Y_i \overset{ind}{\sim} \text{Bernoulli}(p_i)$$

$$log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x_i'}\boldsymbol{\beta}$$

$$\Rightarrow p_i = \frac{e^{\mathbf{x_i'}\boldsymbol{\beta}}}{1+e^{\mathbf{x_i'}\boldsymbol{\beta}}}$$

where $p_i = P(Y_i = 1)$

Model Assumptions

- Linearity of Model
- Independence between observations
- Collinearity does not heavily affect model

Model Selection

Forward Stepwise Selection Algorithm:

---

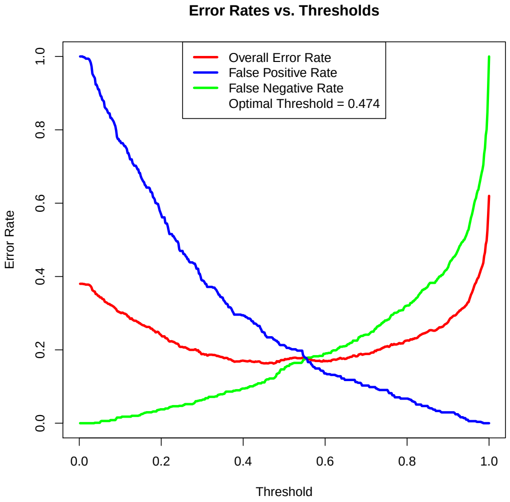
1. Let p be the number of predictors.
2. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
3. For $k = 0, \ldots, p-1$:
   (a) Consider all p-k models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_K + 1$. Here *best* is defined as having smallest RSS or highest R^2.

Results

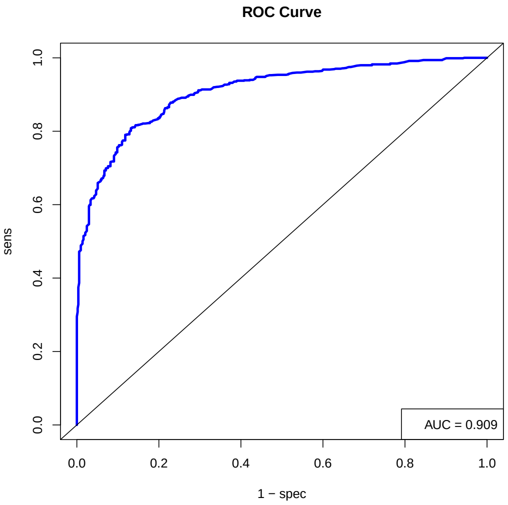
Summary Table

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	98.37	26.64	3.69	0.00
DrugsUnknown	2.53	0.23	11.02	0.00
DrugsYes (drugs involved)	0.64	0.28	2.29	0.02
BeltLap Belt Only Used	-3.02	1.65	-1.83	0.07
BeltNo Helmet	-0.45	1.75	-0.26	0.80
BeltNone Used-Motor Vehicle Occupant	-0.49	1.29	-0.38	0.71
BeltNot Applicable	-2.21	1.73	-1.28	0.20
BeltNot Reported	-3.26	1.72	-1.89	0.06
BeltOther Helmet	0.81	1.35	0.60	0.55
BeltShoulder and Lap Belt Used	-2.75	1.25	-2.20	0.03
BeltShoulder Belt Only Used	-1.79	1.87	-0.96	0.34
BeltUnknown	-2.13	1.30	-1.64	0.10
Speed.RelatedUnknown	2.11	0.51	4.11	0.00
Speed.RelatedYes	1.20	0.19	6.16	0.00
Speed.Limit	0.04	0.01	6.38	0.00
DrinkYes	1.46	0.21	7.14	0.00
LightDark - Not Lighted	0.90	0.23	3.83	0.00
LightDawn	2.21	0.55	4.02	0.00
LightDaylight	1.84	0.24	7.80	0.00
LightDusk	1.08	0.73	1.47	0.14
LightUnknown	0.19	1.50	0.13	0.90
DistractedNot Distracted	1.31	0.31	4.24	0.00
DistractedUnknown	1.67	0.32	5.15	0.00
Mod_year	-0.05	0.01	-3.83	0.00

Thresholds


Error Rates vs. Thresholds
Overall Error Rate
False Positive Rate
False Negative Rate
Optimal Threshold = 0.474
Error Rate
0.0
0.2
0.4
0.6
0.8
1.0
Threshold
0.0
0.2
0.4
0.6
0.8
1.0

Receiver Operating Characteristics (ROC) Curve

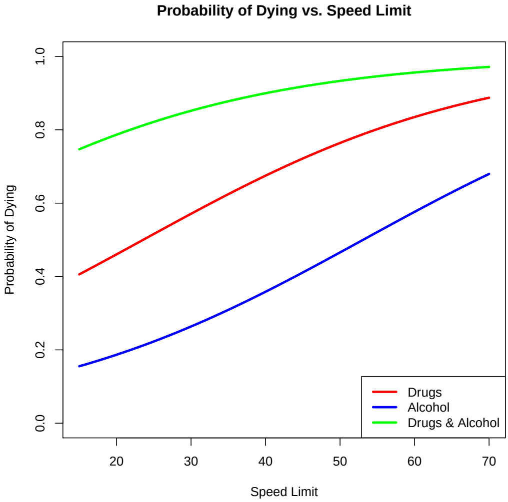


Variance Inflation Factors

VIF Table:

	GVIF	Df	$GVIF^{\frac{1}{2Df}}$
Drugs	1.12	2.00	1.03
Belt	1.24	9.00	1.01
Speed Related	1.16	2.00	1.04
Speed Limit	1.14	1.00	1.07
Drink	1.25	1.00	1.12
Light	1.38	5.00	1.03
Distracted	1.09	2.00	1.02
Model year	1.06	1.00	1.03

Intuition


Probability of Dying vs. Speed Limit
Probability of Dying
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
Speed Limit
Drugs
Alcohol
Drugs & Alcohol

Conclusions

- Model reduced to 8 covariates using forward selection
- AUC $= 91\%$
- VIF < 10

Future

- Look at other combinations of covariates (interaction)