Malik Williams  
4/14/2019  
Data Report Exam 3

1. Introduction

In 1905, the first practical airplane was created. Due to this, there was an increasing interest in improving air travel in the 20th century. After World War I, it showed how effective planes were in transport and as a weapon. However, aviation remained very dangerous because many safety precautions were not invented yet. For example, the only navigation device on most early planes were magnetic compasses. Since these safety precautions were almost nonexistent, fatalities were inevitable. The commercial airline industry was created after passing The Air Mail Act of 1925. The commercialization of air travel prompted a need for safety.[[1]](#footnote-1)

Civil Aviation in the United States is governed by the Federal Aviation Administration (FAA). It was founded in August 1958, replacing the Civil Aeronautics Administration (CAA) and soon after became an agency in the U.S. Department of Transportation. Their role includes developing new forms of aviation technology, setting standards for flight inspections, management of pilot certificates, and creating programs to control the environmental effect of the aviation technology.[[2]](#footnote-2) Because of the FAA, air travel has been optimized in terms of efficiency, safety, and security, making it the most dependable forms of transportation in the world.[[3]](#footnote-3)

The National Transportation Safety Board (NTSB) is a U.S. government agency which investigates civil transportation accidents. It was founded in Washington D.C. and has expanded to have four regional offices located Colorado, Alaska, Virginia and Washington. The NSTB does not only focus on aviation related incidents; it also deals with railroad, marine, highway, and pipeline incidents. They have presented thousands of safety precautions and highlighted the most important safety recommendations. Some safety recommendations for aviation include warning systems that alert the pilot of how close they are to the ground, technology that prevents mid-air collisions, and smoke detectors. Adhering to these suggests significantly decreases the risk of an aviation accident.[[4]](#footnote-4)

2. Data

The dataset below contains data of all aviation incidents occurring in 2016 that were investigated by the NSTB. “Fatal” is a binary variable that is a 1 if the accident had a fatality and 0 otherwise. “Day” is a numerical variable which reports the day of the accident in the year 2016, ranging from 1 to 366. “Lat” is a numerical variable representing the latitude (how many degrees north) of where an accident occurred with negative values representing south. “Lon” is a numerical variable representing the longitude (how many degrees east) of where an accident occurred with negative values representing west. “NonUSA” is a binary variable that is a 0 if the accident took place in the U.S. and is a 1 otherwise. “Phase” is a categorial variable that describes what phase of the flight the accident occurred in; there are many different phases such as maneuvering (MNV), approach (APR), takeoff (TOF), en route (ENR), landing (LDG), initial climb (ICL), etc. (Other). “Event” is a categorical variable that refers to the type of accident that occurred; examples of events are abnormal runway contact (ARC), fuel-related (FUEL), loss of control-ground (LOC-G), system/component failure or malfunction (non-powerplant) (SCF-NP), system/component failure or malfunction (non-powerplant) (SCF-PP), loss of control-inflight (LOC-I), etc. (Other). Personal is a binary variable that is a 1 if the primary purpose of the flight was personal and is 0 otherwise. Multiple is a binary variable that is a 1 if multiple aircrafts were involved in the accident and is 0 otherwise.

Below is a table of the summary statistics of each variable.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fatal | Day | Lat | Lon | NonUSA | Nonairplane | Personal | Multiple |
| Min. | 0 | 1 | -32.84 | -166.71 | 0 | 0 | 0 | 0 |
| 1st Q. | 0 | 126 | 33.45 | -113.77 | 0 | 0 | 0 | 0 |
| Median | 0 | 190 | 38.27 | -94.19 | 0 | 0 | 0 | 0 |
| Mean | .1706 | 190.2 | 38.64 | -98.15 | .01791 | .1215 | .3326 | .02648 |
| 3rd Q. | 0 | 260 | 42.22 | -82.45 | 0 | 0 | 1 | 0 |
| Max. | 1 | 366 | 70.49 | 103.87 | 1 | 1 | 1 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | | Event | |
| LDG | 462 | ARC | 183 |
| ENR | 211 | FUEL | 77 |
| TOF | 153 | LOC-G | 216 |
| MNV | 146 | LOC-I | 215 |
| APR | 121 | Other | 333 |
| ICL | 104 | SCF-NP | 56 |
| Other | 87 | SCF-PP | 204 |

The categorical variables have frequencies of each category as their summary statistics. The binary and numerical variables have their minimum value, 1st quantile, median, mean, 3rd quantile, and maximum value as their summary statistics. Interestingly, based on the mean latitude and longitude, most accidents occur in the northwest (positive latitude, negative longitude). Accidents also seem to rarely occur involving multiple aircrafts.  
  
3. Model Building

I created logistic, probit, and complementary log-log models with “Fatal” as the response variable and all other variables as predictors. Then I performed forward selection and backward elimination on all three models. The resulting models all have the same variables as predictors being Phase, Event, Nonairplane, NonUSA, and Multiple. The AIC values did not change between forward and backward steps. For example, for both forward selection and backward elimination for the logistic regression model, the AIC was 817.2.  
Below is a table containing each model, the stepwise direction, variables in the final models, and the models’ AIC values.

|  |  |  |
| --- | --- | --- |
| **Model** | **Variables** | **AIC** |
| Logistic | Fatal ~ Phase + Event + Nonairplane + NonUSA + Multiple | 817.2 |
| Probit | Fatal ~ Phase + Event + NonUSA + Nonairplane + Multiple | 825 |
| C. Log-Log | Fatal ~ Phase + Event + Nonairplane + Multiple + NonUSA | **807.5** |

4. Model Assessment

Based on AIC values, the complementary log-log regression model with Phase, Event, Nonairplane, NonUSA, and Multiple as predictors produced the lowest AIC value (807.5). Therefore, it is the best model.

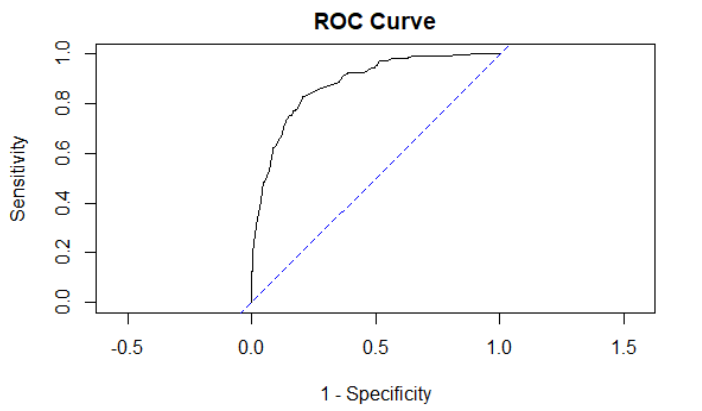
**Correlation Measure R = 0.5875619  
Likelihood Measure D = 0.3422826**

|  |  |  |  |
| --- | --- | --- | --- |
| **Apparent Classification Table** | | **yhat** | |
| 0 | 1 |
| **y** | 0 | 1013 | 52 |
| 1 | 113 | 106 |

**Estimated Sensitivity: 0.4840183  
Estimated Specificity: 0.9511737**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cross-Validated Classification Table** | | **yhat** | |
| 0 | 1 |
| **y** | 0 | 1012 | 53 |
| 1 | 113 | 106 |

**Cross-validated Sensitivity: 0.4840183  
Cross-validated Specificity: 0.9502347**

  
**Concordance Index: 0.8796493**

For prediction, this model is very good at predicting non-fatal accidents because its specificity is .95. However, when predicting for fatal accidents, it is not good at predicting fatal accidents because its sensitivity is .48. The sensitivity seems to be equivalent to guessing. Therefore, it is not an effective model because it does not reliably predict fatal accidents (sensitivity is too low).

5. Conclusion

In conclusion, when predicting if an accident is fatal, knowing the phase the aircraft was in, the event that caused the accident, whether or not the aircraft was an airplane, whether or not the accident occurred in the United States or not, and if multiple aircrafts were involved with the accident are all good predictors. Using the log-log regression model produced by the step function, given that there is not a fatal accident, there is 95% chance that the model correctly predicts it. Given that there is a fatal accident, there is 48% chance that the model correctly predicts it. This does not have significant difference than if we randomized it. We would expect a 50% chance of correctly predicting a fatal accident since the response is binary (either fatal or nonfatal). This could be because the NSTB is more focused on establishing preventive measures.

6. Appendix

**Load in Dataset & Producing Summary Statistics**  
accidents = read.csv("NTSBGA2016.csv")  
summary(accidents)

**Creating the 3 Regression Models**  
mod.log <- glm(Fatal ~ ., data = accidents, family = binomial) #Logistic Model  
mod.probit <- glm(Fatal ~ ., family = binomial(link = "probit"), data = accidents) #Probit Model  
mod.cloglog <- glm(Fatal ~ ., family = binomial(link = "cloglog"), data = accidents) #Comp. Log-Log Model

**Forward Selection and Backward Elimination on the Logistic Model**  
nullmodlog <- glm(Fatal ~ 1, family=binomial, data=accidents)  
step(nullmodlog,~ Day + Lat+ Lon + NonUSA + Nonairplane + Phase + Event + Personal + Multiple, direction="forward") #Forward  
step(mod.log) #Backward

**Forward Selection and Backward Elimination on the Probit Model**  
nullmodprobit <- glm(Fatal ~ 1, family=binomial(link="probit"), data=accidents)  
step(nullmodprobit,~ Day + Lat+ Lon + NonUSA + Nonairplane + Phase + Event + Personal + Multiple, direction="forward") #Forward  
step(mod.probit) #Backward

**Forward Selection and Backward Elimination on the Complementary Log-log Model**  
nullmodloglog <- glm(Fatal ~ 1, family=binomial(link="cloglog"), data=accidents)  
step(nullmodloglog,~ Day + Lat+ Lon + NonUSA + Nonairplane + Phase + Event + Personal + Multiple, direction="forward") #Forward  
step(mod.cloglog) #Backward

**Creating the Best Model (based on AIC values)**  
best\_mod <- glm(Fatal ~ NonUSA + Nonairplane + Phase + Event + Multiple, family = binomial(link "cloglog"), data = accidents)

**Correlation Measure R**  
cor(accidents$Fatal, fitted(best\_mod)) #correlation measure

**Likelihood Measure D**  
as.numeric((logLik(best\_mod) - logLik(nullmodloglog)) / (0 - logLik(nullmodloglog)))

**Apparent Classification Table**  
pi0 <- 0.5  
table(y=accidents$Fatal, yhat=as.numeric(fitted(best\_mod) > pi0))

**Sensitivity and Specificity Calculation**  
106 / (106 + 113) # Sensitivity  
1013 / (1013 + 52) # Specificity

**Cross-Validation Table**   
pihatcv <- numeric(nrow(accidents))  
for(i in 1:nrow(accidents)){  
 pihatcv[i] <- predict(update(best\_mod, subset=-i), newdata=accidents[i,], type="response")}  
table(y=accidents$Fatal, yhat=as.numeric(pihatcv > pi0))

**Sensitivity and Specificity Calculation**  
106 / (106 + 113) # Sensitivity  
1012 / (1012 + 53) # Specificity

**Plotting ROC Curve**  
pihat <- fitted(best\_mod)  
false.neg <- c(0,cumsum(tapply(accidents$Fatal,pihat,sum)))  
true.neg <- c(0,cumsum(table(pihat))) - false.neg  
plot(1-true.neg/max(true.neg), 1-false.neg/max(false.neg), type="l", main="ROC Curve", xlab="1 - Specificity", ylab="Sensitivity", asp=1)  
abline(a=0, b=1, lty=2, col="blue")

**Concordance Index**  
mean(outer(pihat[accidents$Fatal==1], pihat[accidents$Fatal==0], ">") + 0.5 \* outer(pihat[accidents$Fatal==1], pihat[accidents$Fatal==0], "=="))

1. <https://www.faa.gov/about/history/brief_history/> [↑](#footnote-ref-1)
2. <https://en.wikipedia.org/wiki/Federal_Aviation_Administration> [↑](#footnote-ref-2)
3. <https://www.faa.gov/about/history/brief_history/#origins> [↑](#footnote-ref-3)
4. <https://en.wikipedia.org/wiki/National_Transportation_Safety_Board> [↑](#footnote-ref-4)