MA 576 HW 3

Benjamin Draves

4

```
#load necessary packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

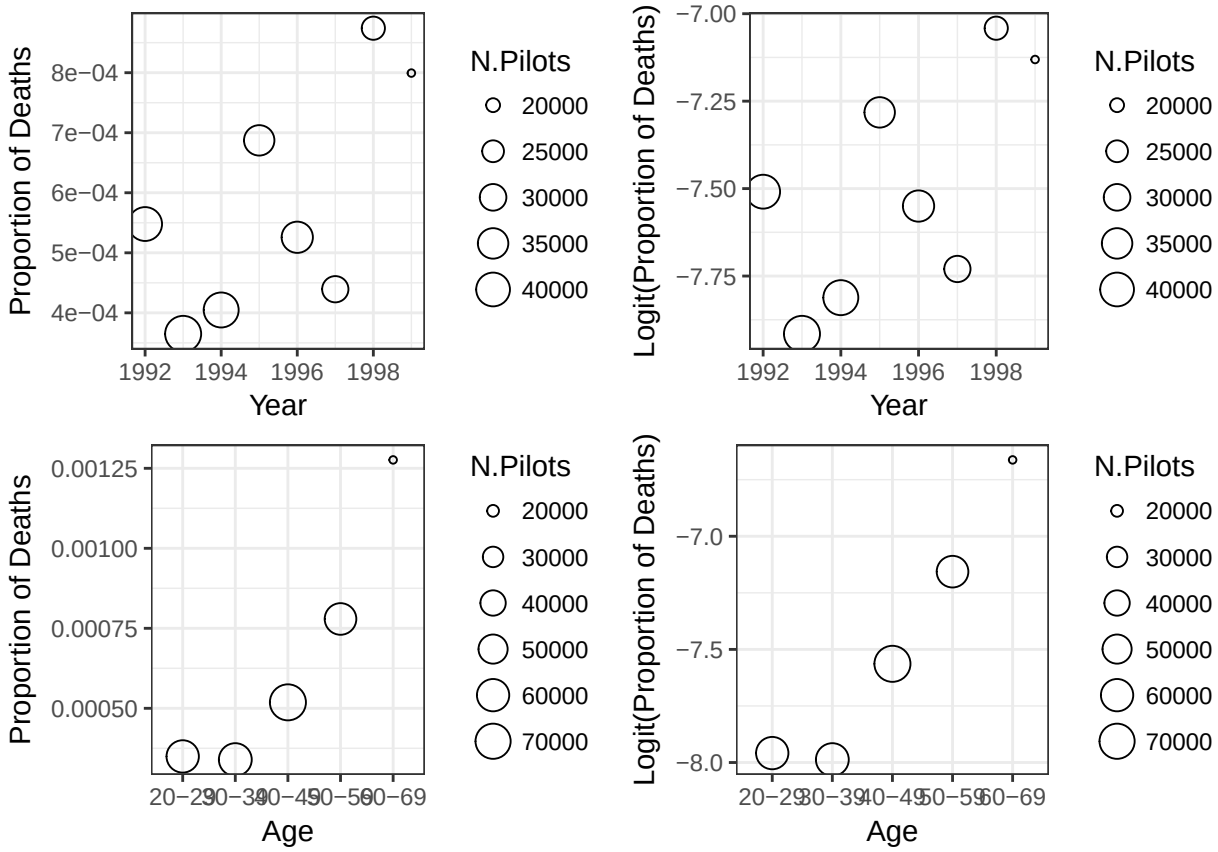
```
library(tidyr)
library(ggplot2)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

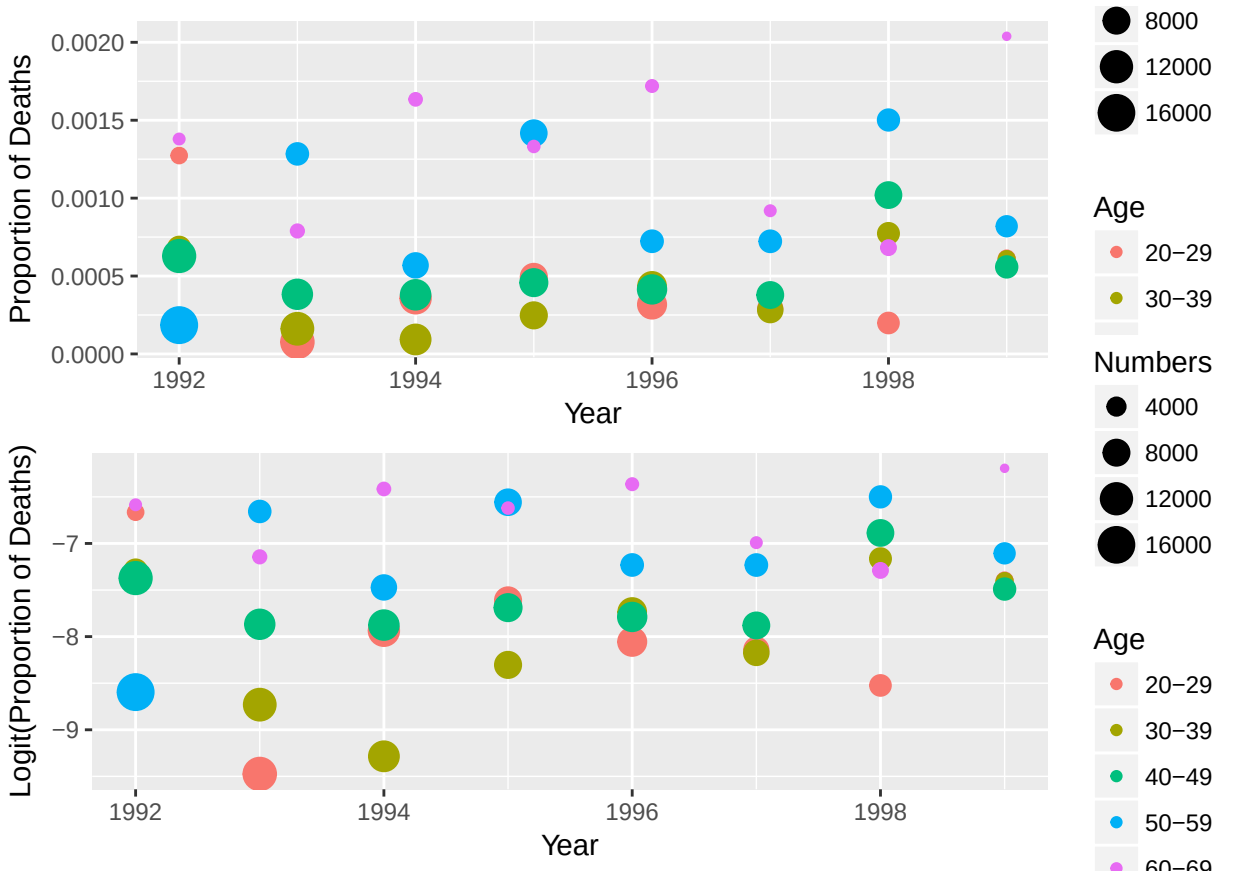
```
#read in data
planes = read.table("~/Desktop/Courses/MA 576/data/aviation.txt", header = T)

#define logistic function
logit = function(p) log(p/(1-p))

#Make proportions/tables
planes.yr = planes %>%
            group_by(Year) %>%
            summarise(N.Pilots = sum(Numbers), N.Deaths = sum(Deaths))%>%
            mutate(prop.deaths = N.Deaths/N.Pilots,
                   logit.prop.deaths = logit(prop.deaths))

planes.ag = planes %>%
            group_by(Age) %>%
            summarise(N.Pilots = sum(Numbers), N.Deaths = sum(Deaths))%>%
            mutate(prop.deaths = N.Deaths/N.Pilots,
                   logit.prop.deaths = logit(prop.deaths))
```

```r
#visualize death % by year
p1 <- ggplot(planes.yr, aes(x = Year, y = prop.deaths, size = N.Pilots)) +
      labs(y = "Proportion of Deaths")+
      theme_bw()
p2 <- ggplot(planes.yr, aes(x = Year, y = logit.prop.deaths, size = N.Pilots)) +
      labs(y = "Logit(Proportion of Deaths)")+
      theme_bw()

#visualize death % by age
p3 <- ggplot(planes.ag, aes(x = Age, y = prop.deaths, size = N.Pilots)) +
      labs(y = "Proportion of Deaths")+
      theme_bw()
p4 <- ggplot(planes.ag, aes(x = Age, y = logit.prop.deaths, size = N.Pilots)) +
      labs(y = "Logit(Proportion of Deaths)")+
      theme_bw()

grid.arrange(p1, p2, p3, p4, nrow=2, ncol = 2)
```




```r
#visualize the effect of age and year simultaneously
planes$prop.deaths = planes$Deaths/planes$Numbers
planes$logit.prop.deaths = logit(planes$prop.death)

p5 = ggplot(planes, aes(x = Year, y = prop.deaths,color = Age, size = Numbers)) +
```

```
    geom_point()+
    labs(y = "Proportion of Deaths")
p6 = ggplot(planes, aes(x = Year, y = logit.prop.deaths,
                        color = Age,size = Numbers)) +
    geom_point()+
    labs(y = "Logit(Proportion of Deaths)")
grid.arrange(p5,p6, nrow = 2)
```



Above are six plots. The first four plots are marginal bubble plots of the proportion of deaths against year and age and the logistic proportion of deaths against year age. The size of each bubble corresponds to the number of registered pilots in that year or age group. We notice in general, there is a slight increase in the proportion of aviation deaths over time, but this trend could be attributable to noise. The proportion of deaths against age, however, exhibits a clear positive correlation between age and proportion of deaths. That is there are more aviation deaths for older pilots compared to their younger counterparts. In general, the logistic mappings provide the same insights as the untransformed data.

The last two plots attempt to visualize this trend simultaneously. The age groups are colored as the x axis is time is plotted against the response variable. There appears to be little variance explained by year but the age groups are almost always ordered by age on the y-axis.

b

```r
m0 = glm(prop.deaths ~1, data = planes, family = binomial, weights = Numbers)
m1 = glm(prop.deaths ~ Age + as.factor(Year), data = planes,
         family = binomial, weights = Numbers)
summary(m1)
```

```
##
## Call:
## glm(formula = prop.deaths ~ Age + as.factor(Year), family = binomial,
##     data = planes, weights = Numbers)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.87803  -0.40183  -0.00675   0.49640   2.26381
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -8.05724    0.31422 -25.642  < 2e-16 ***
## Age30-39            -0.02898    0.30882  -0.094  0.92522
## Age40-49             0.36531    0.27383   1.334  0.18218
## Age50-59             0.76586    0.27008   2.836  0.00457 **
## Age60-69             1.22258    0.30123   4.059 4.93e-05 ***
## as.factor(Year)1993 -0.17915    0.33535  -0.534  0.59320
## as.factor(Year)1994 -0.11623    0.32754  -0.355  0.72269
## as.factor(Year)1995  0.36047    0.29815   1.209  0.22665
## as.factor(Year)1996  0.14207    0.31833   0.446  0.65539
## as.factor(Year)1997 -0.07978    0.35322  -0.226  0.82130
## as.factor(Year)1998  0.53002    0.30216   1.754  0.07942 .
## as.factor(Year)1999  0.41304    0.33697   1.226  0.22030
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.418  on 39  degrees of freedom
## Residual deviance: 36.246  on 28  degrees of freedom
## AIC: 183.08
##
## Number of Fisher Scoring iterations: 5
```

Here we construct a binomial glm with predictors Age and Year which are considered categorical variables. As a result, there are 12 covariates. We choose to interpret a single age variable, a single year variable, as well as the intercept in this model.

The intercept for this model is given by $\hat{\beta}_0 = -8.05724$. This value can be interpreted as follows; the odds a pilot in the age range 20 - 29 in the year 1992 with die in an aviation accident is $e^{-8.05724} = 0.0003168$. Identically a pilot in the age range 20 - 29 in the year 1992 with die in an aviation accident with probability 0.0003166996.

The coefficient for the $60 - 69$ age group is given by $\hat{\beta}_{60-69} = 1.22258$. This value can be

interpreted as follows; the odds a pilot in the age range 60-69 in 1992 dies in an aviation accident is $e^{1.22258} = 3.395938$ greater than a pilot in the age range $20 - 29$. This corresponds to a probability of death given by

$$p_{death} = \frac{\exp -8.05724 + 1.22258}{1 + \exp -8.05724 + 1.22258} = 0.001074677$$

The coefficient for the year 1998 is given by $\hat{\beta}_{1998} = 0.53002$. This value can be interpreted as follows; the odds a pilot in the age range 20-29 in 1998 dies in an aviation accident is $e^{0.53002} = 1.698966$ greater than a pilot in the age range $20 - 29$ in 1992. This corresponds to a probability of death given by

$$p_{death} = \frac{\exp -8.05724 + 0.53002}{1 + \exp -8.05724 + 0.53002} = 0.0005379429$$

The only significant variables in this model are the intercept and age groups $50 - 59$ and $60 - 69$.

(c)

```
m2 = glm(prop.deaths ~ Age, data = planes,
         family = binomial, weights = Numbers)
summary(m2)
```

```
##
## Call:
## glm(formula = prop.deaths ~ Age, family = binomial, data = planes,
##     weights = Numbers)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2658  -0.5895  -0.0547   0.6736   2.1307
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.95826    0.21826 -36.463  < 2e-16 ***
## Age30-39    -0.02886    0.30866  -0.093  0.92551
## Age40-49     0.39447    0.27197   1.450  0.14693
## Age50-59     0.80175    0.26434   3.033  0.00242 **
## Age60-69     1.29571    0.29892   4.335 1.46e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.418  on 39  degrees of freedom
## Residual deviance: 45.495  on 35  degrees of freedom
## AIC: 178.33
##
## Number of Fisher Scoring iterations: 5
```

By removing the year covariant, we see slight changes to the parameter estimates. In most every

case, save the intercept we see a slight increase in coefficients suggesting that there is variance in aviation deaths that is explained by both age and year.

```
anova(m0, m2, m1, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: prop.deaths ~ 1
## Model 2: prop.deaths ~ Age
## Model 3: prop.deaths ~ Age + as.factor(Year)
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        39     74.418
## 2        35     45.495  4  28.9229 8.104e-06 ***
## 3        28     36.246  7   9.2487    0.2353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After constructing an analysis of deviance table, we see that the model with age explains a significant amount of variances in the response while the addition of year proved insignificant. We do note however, that in some of our models, there could be slight overdispersion which affect these test statistics.

(d)

```
m3 = glm(prop.deaths ~as.numeric(Age), data = planes,
         family = binomial, weights = Numbers)
summary(m3)
```

```
##
## Call:
## glm(formula = prop.deaths ~ as.numeric(Age), family = binomial,
##     data = planes, weights = Numbers)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3146  -0.6507   0.0659   0.8450   2.4297
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.53229    0.23301 -36.618  < 2e-16 ***
## as.numeric(Age)  0.34807    0.06814   5.108 3.25e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 74.418  on 39  degrees of freedom
## Residual deviance: 47.614  on 38  degrees of freedom
## AIC: 174.45
```

```
##
## Number of Fisher Scoring iterations: 5
```

The intercept estimate in this model $\hat{\beta}_0 = -8.53229$ can be interpreted as the odds that a "zero year old pilot" dies in a plane accident is given by $e^{-8.53229}$. That is, $\hat{\beta}_0$ has little practical meaning in this context.

Now notice that in this context we model

$$\log \frac{p_{death}}{1 - p_{death}} = \beta_0 + \beta_1 Age$$

Rewriting in terms of odds we have

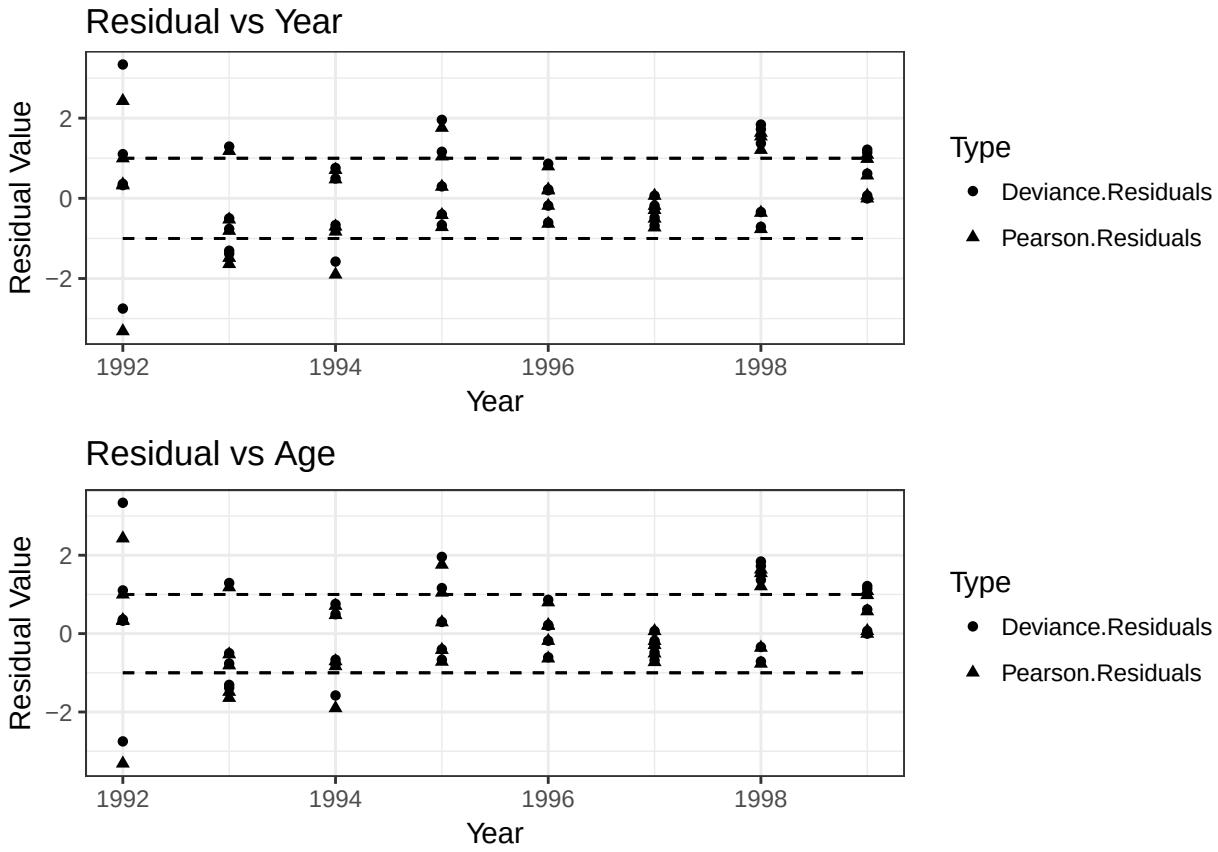
$$Odds = e^{\beta_0}(e^{\beta_1})^x$$

Therefore, we can interpret $\hat{\beta}_1 = 0.34807$ as follows; the odds a pilot dies in an aviation accident is $e^{0.34807} = 1.416331$ times more likely each year. This interpretation is different from that in (c) as it is recursive and not with respect to some reference group.

The pros of this model is that it is more parsimonious in the sense that less covariates are used. This allows for inference to occur through a single parameter instead of several parameters as before. One possible drawback to this framework is that the interpretation of the parameters are in terms of previous odds and not to a specific reference group. By treating age as a factor, we are able to reference a single group's odds instead of recursively as we do in this model.

(e)

```
dev.res = resid(m3,type='pearson')
pear.res = resid(m3,type='deviance')

Residuals = data.frame(Age = planes$Age, Year = planes$Year,
                       Deviance.Residuals = dev.res,
                       Pearson.Residuals = pear.res)
res = Residuals %>% gather("Type", "Residual.Value", 3:4)

p7 = ggplot(res, aes(x = Year, y = Residual.Value, shape = Type)) +
  geom_point()+
  labs(y = "Residual Value", title = "Residual vs Year")+
  geom_line(aes(y = 1), linetype = 2)+
  geom_line(aes(y = -1), linetype = 2)+
  theme_bw()
p8 = ggplot(res, aes(x = Year, y = Residual.Value, shape = Type)) +
  geom_point()+
  labs(y = "Residual Value", title = "Residual vs Age")+
  geom_line(aes(y = 1), linetype = 2)+
  geom_line(aes(y = -1), linetype = 2)+
  theme_bw()


grid.arrange(p7,p8, nrow = 2)
```



Above are two plots of the Pearson and Deviance residuals against year and age. In the age only model, we estimate the residual deviance as 47.614 on 39 degrees of freedom. Normally we anticipate that each residual is less than 1 but this estimate suggests that these values could be larger for this model. We see this behavior in both deviance and Pearson residuals (we expect this behavior as show in problem 1). In general, while we may have some overdispersion issues, this model seems quite appropriate.