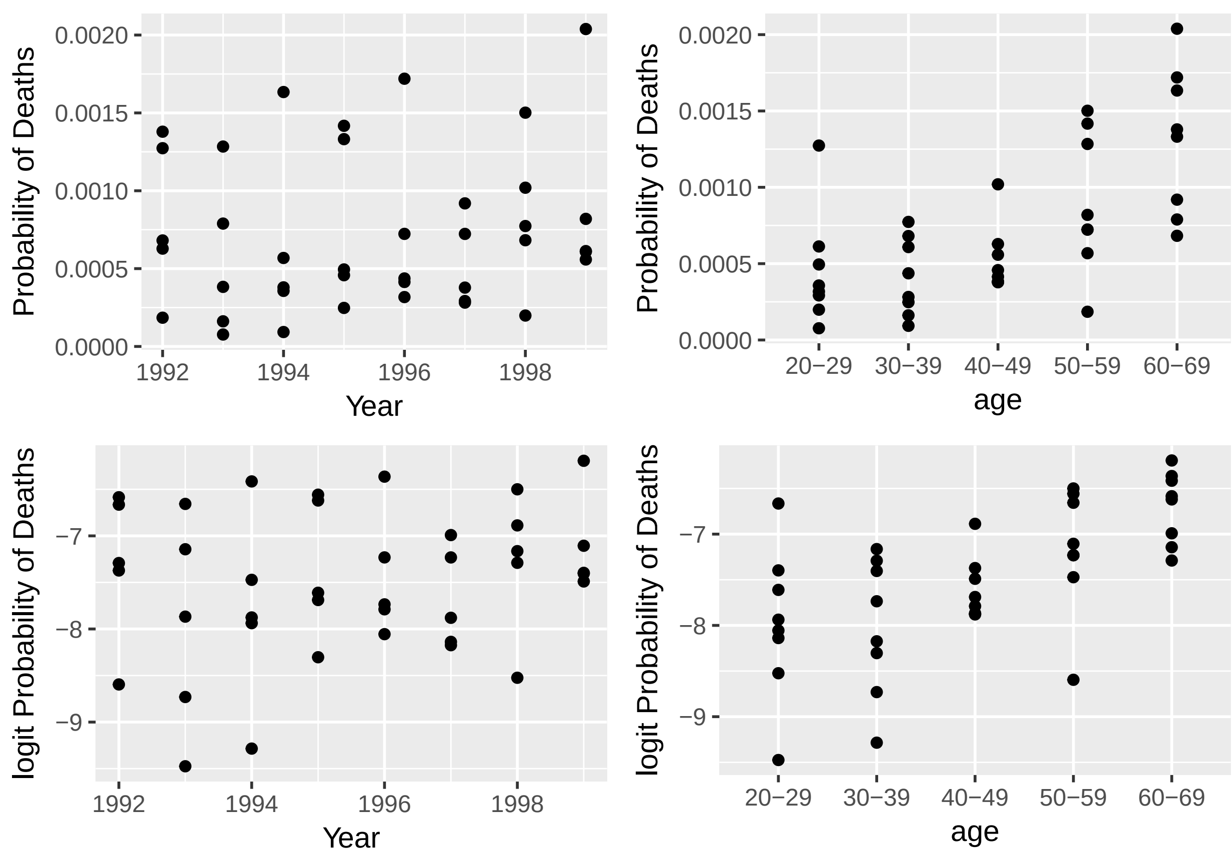
MA576 HW3

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```r
##4
#a
aviation <- read.delim("aviationdeaths.txt", header = T)
attach(aviation)
yf <- as.factor(Year)
library(ggplot2)
library(grid)
library(gridExtra)
proportion <- Deaths/Numbers
logitp <- log(proportion/(1 - proportion))
p1 <- qplot(Year, proportion, xlab = "Year", ylab = "Probability of Deaths")
p2 <- qplot(Age, proportion, xlab = "age", ylab = "Probability of Deaths")
p3 <- qplot(Year, logitp, xlab = "Year", ylab = "logit Probability of Deaths")
p4 <- qplot(Age, logitp, xlab = "age", ylab = "logit Probability of Deaths")
grid.arrange(p1, p2, p3, p4, nrow = 2, ncol = 2)
```



The probabilities of accidents-caused deaths is very small (three-decimal). The logit of proportions becomes quite large in the absolute term, which also indicated that probabilities of accidents-caused death is very small. It seems that both of them distributed randomly for Year groups. However, both of them tend to increase a age increases.

```r
#b
mod1 <- glm(cbind(Deaths, Numbers - Deaths) ~ Age +yf, family = binomial(link ='logit'), data = aviatio
summary(mod1)
```

```
##
## Call:
## glm(formula = cbind(Deaths, Numbers - Deaths) ~ Age + yf, family = binomial(link = "logit"),
##     data = aviation)
##
## 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 ***
## yf1993      -0.17915    0.33535  -0.534  0.59320
## yf1994      -0.11623    0.32754  -0.355  0.72269
## yf1995       0.36047    0.29815   1.209  0.22665
## yf1996       0.14207    0.31833   0.446  0.65539
## yf1997      -0.07978    0.35322  -0.226  0.82130
## yf1998       0.53002    0.30216   1.754  0.07942 .
## yf1999       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
```

The intercept term (b0) indicates the log odds of age "20-29" have deaths in year 1992. b1 for Age30-39 means the log odds ratio between age "20-29" in year 1992 and age30-39 in year 1992 b2 for Age40-49 means the log odds ratio between age "20-29" in year 1992 and age40-49 in year 1992 b3 for Age50-59 means the log odds ratio between age "20-29" in year 1992 and age50-59 in year 1992 b4 for Age60-69 means the log odds ratio between age "20-29" in year 1992 and age60-69 in year 1992 b5 for Year1993 means the log odds ratio between age "20-29" in year 1992 and age "20-29" in year 1993 b6 for Year1994 means the log odds ratio between age "20-29" in year 1992 and age "20-29" in year 1994 b7 for Year1995 means the log odds ratio between age "20-29" in year 1992 and age "20-29" in year 1995 b8 for Year1996 means the log odds ratio between age "20-29" in year 1992 and age "20-29" in year 1996 b9 for Year1997 means the log odds ratio between age "20-29" in year 1992 and age "20-29" in year 1997 b10 for Year1998 means the log odds ratio between age "20-29" in year 1992 and age "20-29" in year 1998 b11 for Year1999 means the log odds ratio between age "20-29" in year 1992 and age "20-29" in year 1999 The intercept and coeefiecients of age50-59 and age60-69 groups seem to have significance in predicting of the model.

```r
#c
mod2 <- glm(cbind(Deaths, Numbers - Deaths) ~ Age, family = binomial(link ='logit'), data = aviation)
summary(mod2)
```

```
##
## Call:
## glm(formula = cbind(Deaths, Numbers - Deaths) ~ Age, family = binomial(link = "logit"),
##     data = aviation)
##
## 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
```

The intercept term (b0) indicates the log odds of age "20-29" between have deaths and not. b1 for Age30-39 means the log odds ratio between age "20-29" and age30-39 b2 for Age40-49 means the log odds ratio between age "20-29" and age40-49 b3 for Age50-59 means the log odds ratio between age "20-29" and age50-59 b4 for Age60-69 means the log odds ratio between age "20-29" and age60-69 the significant parameter estimates are still intercept term, age50-59 and age60-69 without the factor of years. P-value is even smaller than the model in part b.

```r
anova(mod1,test = "Chisq") #with both factors age and year
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Deaths, Numbers - Deaths)
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    39     74.418
## Age   4  28.9229        35     45.495 8.104e-06 ***
## yf    7   9.2487        28     36.246    0.2353
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod2, test = "Chisq") # with only the factor of age
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Deaths, Numbers - Deaths)
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    39     74.418
## Age   4   28.923        35     45.495 8.104e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mod1, mod2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(Deaths, Numbers - Deaths) ~ Age + yf
## Model 2: cbind(Deaths, Numbers - Deaths) ~ Age
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        28     36.246
## 2        35     45.495 -7  -9.2487   0.2353
```

As we can see from three analysis of deciance table, when comparing to the null model. Age seems to always be significant. When comparing nested model and restricted model, year seems doesn't improve the model significantly.

```r
#d
agen <- as.numeric(Age)
mod3 <- glm(cbind(Deaths, Numbers - Deaths) ~ agen, family = binomial(link ='logit'), data = aviation)
summary(mod3)
```

```
##
## Call:
## glm(formula = cbind(Deaths, Numbers - Deaths) ~ agen, family = binomial(link = "logit"),
##     data = aviation)
##
## 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 ***
## agen         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
```

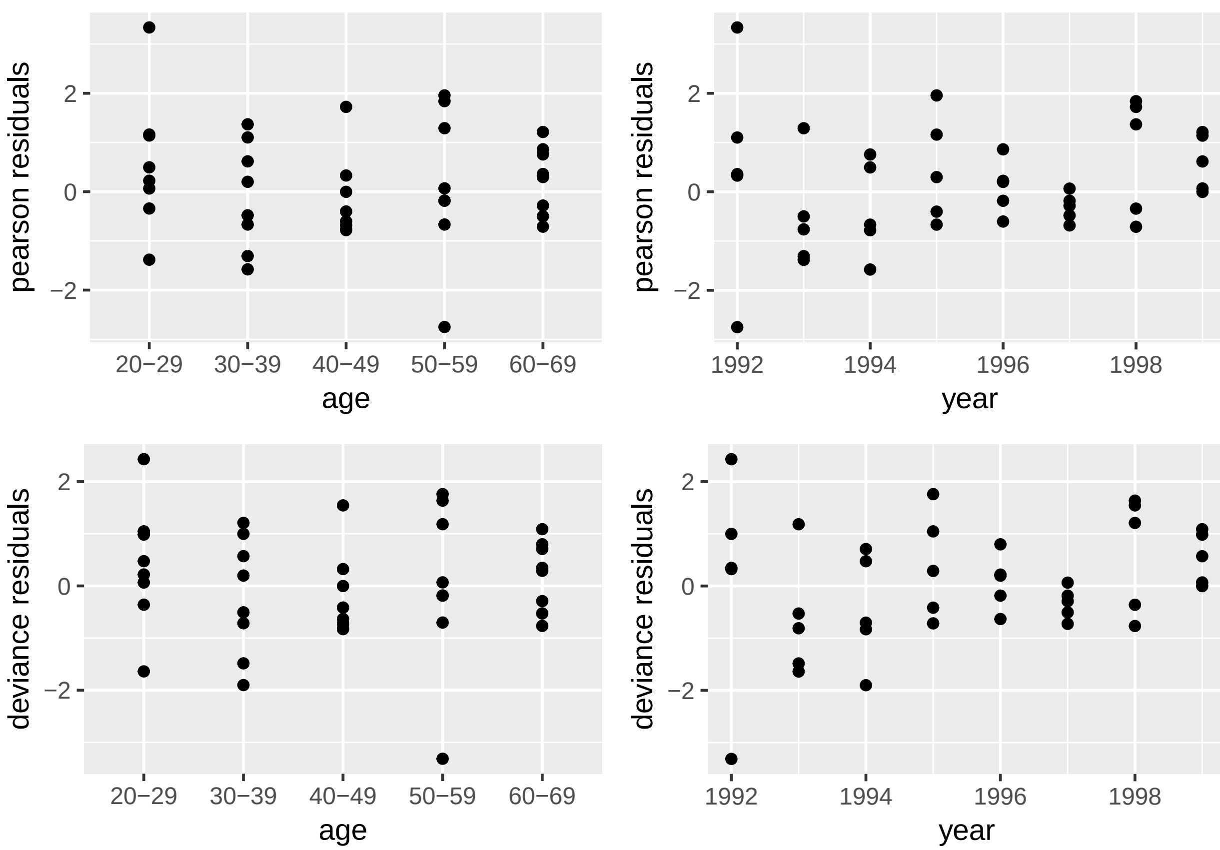
The intercept term means the log odds of deaths if age is 0 (not make sense in this model). Coefficient of age (b1) means the log odds ratio of deaths when X increase by 1. For example, probability of deaths of age 20 is 0.0001970033 * (1.416331^20) = 0.208. The intercept and age terms are both significant. Advantages of the factor models is that it tells us wich age group has significant effect on deaths counts yet when age is numeric, we assume each unit increase of age is significant. Advantages of the numeric model is that we could have prediction of each age if we need, yet the factor model is limited to predict with the accuracy to each age.

I think the factor model provides the most parismonious model since we don't need data of deaths count for each age, but only five groups of age could be enough. It is also effectively tells us that age 20-49 does not have significant effect on the probability of deaths.

```r
#e
pres <- residuals(mod3, type = "pearson")
dres <- residuals(mod3, type = "deviance")
dispersion <- sum((pres^2)/39) #n-p = 40-1 = 39
deviance(mod3)
```

```
## [1] 47.61432
```

```r
ppres1 <- qplot(Age, pres, xlab = "age", ylab = "pearson residuals")
ppres2 <- qplot(Year, pres, xlab = "year", ylab = "pearson residuals")
pdres1 <- qplot(Age, dres, xlab = "age", ylab = "deviance residuals")
pdres2 <- qplot(Year, dres, xlab = "year", ylab = "deviance residuals")
grid.arrange(ppres1, ppres2, pdres1, pdres2, nrow = 2, ncol = 2)
```



the level of dispersion is 1.29683. Residuals seems fit randomly around zero as year or age increases. AIC is less than the factor model and factor model with factor year.

```r
anova(mod3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Deaths, Numbers - Deaths)
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    39     74.418
## agen  1   26.804        38     47.614 2.252e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The analysis of deviance table shows that age this is a significant model.