Homework 4

Tyler Poelking

2/21/2017

1.a

```
hiro <- read_delim("http://www.stat.osu.edu/~pfc/teaching/3302/datasets/hiroshima.txt",
    " ", escape_double = FALSE, trim_ws = TRUE)
```

```
## Parsed with column specification:
## cols(
##   radiation = col_character(),
##   midpoint = col_double(),
##   leukemia = col_integer(),
##   other = col_integer(),
##   total = col_integer()
## )
```

```
## Warning: 1 parsing failure.
## row col  expected    actual
##   6  -- 5 columns 6 columns
```

```
rad.l <-
  factor(ifelse(hiro$radiation=="0", "0",
                ifelse(hiro$radiation=="1to9", "1to9",
                       ifelse(hiro$radiation=="10to49", "10to49",
                              ifelse(hiro$radiation=="50to99", "50to99",
                                     ifelse(hiro$radiation=="100to199", "100to199", "200plus" ))))),
         levels=c("0", "1to9","10to49","50to99","100to199","200plus"))
```

1.b

```
options(contrasts=c("contr.treatment", "contr.poly"))
p = hiro$leukemia/hiro$total
hiro = cbind(p,hiro)
hiro.logit.l =  glm(p ~ rad.l, weight = total, family = binomial, data=hiro)

summary(hiro.logit.l)
```

```
##
## Call:
## glm(formula = p ~ rad.l, family = binomial, data = hiro, weights = total)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.3699     0.2821 -11.947  < 2e-16 ***
## rad.l1to9      -0.3189     0.5334  -0.598   0.5499
## rad.l10to49    -0.0379     0.5350  -0.071   0.9435
## rad.l50to99     0.6184     0.6589   0.939   0.3480
## rad.l100to199   1.3222     0.6015   2.198   0.0279 *
## rad.l200plus    2.7638     0.4067   6.795 1.08e-11 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5.4351e+01  on 5  degrees of freedom
## Residual deviance: 1.3323e-14  on 0  degrees of freedom
## AIC: 33.665
##
## Number of Fisher Scoring iterations: 4
```

```
bo =  c((-3.3699 - 1.96*0.2821),(-3.3699 + 1.96*0.2821 ))

CI1to9 = c((-.3189 - 1.96*0.5334),(-.3189 + 1.96*0.5334 ))

CI10to49 = c((-0.0379 - 1.96*0.5350), (-0.0379 + 1.96*0.5350))

CI50to99 = c((0.6184 - 1.96* 0.6589), (0.6184 + 1.96* 0.6589))

CI100to199 = c((1.322 - 1.96*0.6015), (1.322 + 1.96*0.6015))

CI200plus = c((2.7638 - 1.96*0.4067),(2.7638 + 1.96*0.4067))
```

According to our model, the intercept is significantly different than 0. The coefficients corresponding to a radiation levels 100to199 and 200plus are also significantly different than zero, indicating that the luekemia rate of dead cancer patients that received radation in 100to199 and 200plus are significantly different than the luekemia rate of dead cancer patients who received radiation in the 0 level. Coefficients corresponding to levels 1to9, 10to49, and 50to99 are not significantly different from zero, indicating patients in these levels do not have luekemia rates different than patients with those who received a radiation level of 0.

Fixing color, the estimated odds of having a dead patient multiplies by e^1.3222 = 3.751666 for a patient in radiation level 100to199, relative to a patient with 0 radiation level. In other words, it is how much more likely a patient with 100to199 radiation is expected to have luekemia compared to a patient with 0 level radiation. A 95% confidence interval for this value is (1.153799, 12.19395).

Fixing color, the estimated odds of having a dead patient multiplies by e^2.7638 = 15.86 for a patient in radiation level 200plus, relative to a patient with 0 radiation level. In other words, it is how much more likely a patient with 200plus radiation is expected to have luekemia compared to a patient with 0 level radiation. A 95% confidence interval for this value is (7.146824, 35.19598).

The other factor variables were not significant so they were not mentioned, but C.I.'s for their odds ratios would be calculated similarly to the two above.

95% Confidence intervals for the coefficients corresponding to these factor variables representing age groups are as follows.

Intercept(B0): [-3.922816, -2.816984]

CI1to9: [-1.364364, 0.726564]

CI10to49: [-1.0865, 1.0107]

CI50to99: [-0.673044, 1.909844]

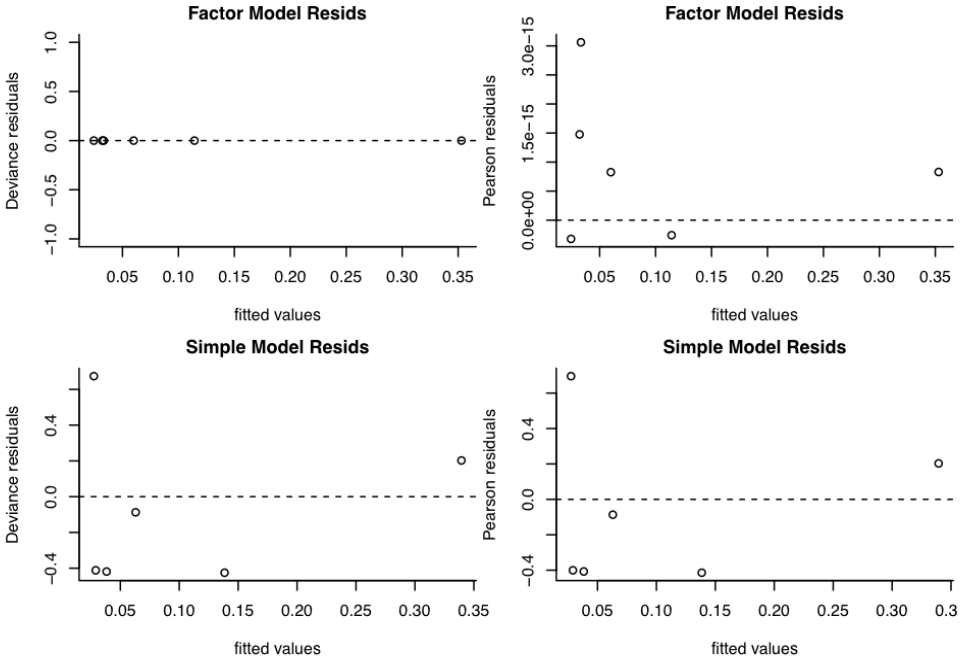
CI100to199: [0.14306, 2.50094]

CI200plus: [1.966668, 3.560932]

1.c

##

```
## Call:
## glm(formula = p ~ midpoint, family = binomial, data = hiro, weights = total)
##
## Deviance Residuals:
##        1         2         3         4         5         6
##  0.67399  -0.41184  -0.41877  -0.08743  -0.42526   0.20237
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.565875   0.212254 -16.800  < 2e-16 ***
## midpoint     0.011624   0.001487   7.819 5.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 54.3509  on 5  degrees of freedom
## Residual deviance:  1.0287  on 4  degrees of freedom
## AIC: 26.694
##
## Number of Fisher Scoring iterations: 4
```



The AIC for the simple logistic regression model used in Homework 3 is 26.694 while the factor model above's is 33.665. So based on the AIC, the preffered model is simple logistic regression model. The residual plots have fairly similar spreads to one another, with the exception of the deviance residual plot of the factor model. The points on this plot fall exactly in line with zero. This is due to the fact that the fitted values mapped directly to the sample proportions. Both simple logistic regression model resid plots look the same, and look fairly similar to the pearson resid plot for the factor model, being more or less centered around the

0 and more points fitted to lower fitted values (more points on the left side) with one point far out at a fitted value of .3. From these plots, I would favor the Simple Model because the deviance resid plot for the factor model gives way to potential overfit. The residual deviances for the factor model are practically zero on zero degrees of freedom. This means that the model fits very close to the saturated model, which is not good due to overfit. It in in this sense that the simple logistic regression model is preffered, because it has a great residual deviance than 0 but it is still low (1.0287 on 4 degrees of freedom)

1.d The coefficient for 50to90 is 0.6184 and the coefficient for 100to199 is 1.3222. The difference is 1.3222-0.6184=0.7038, indicating that the ratio change when going from 100to199 to 50to99 is e^0.7038 = 2.021622. In other words, a patients with radiation at 100to199 are 2.021622 more likely to have been a luekemia patient. This ratio can also be found by a transpose multiplied by the B-hat vector. The contents of a transpose = (0, 0 , 0 , -1, 1, 0) and the content of B-hat = (-3.3699, -0.3189, -0.0379, 0.6184, 1.3222, 2.7638). Taking the e to the multiple of these two vectors produces 2.0218.

```
a = c(0, 0, 0, -1, 1, 0)
beta.hat = round(coef(hiro.logit.l), 3)
est.log.odds = t(a) %*% beta.hat
est.odds = round(exp(est.log.odds), 4)
```

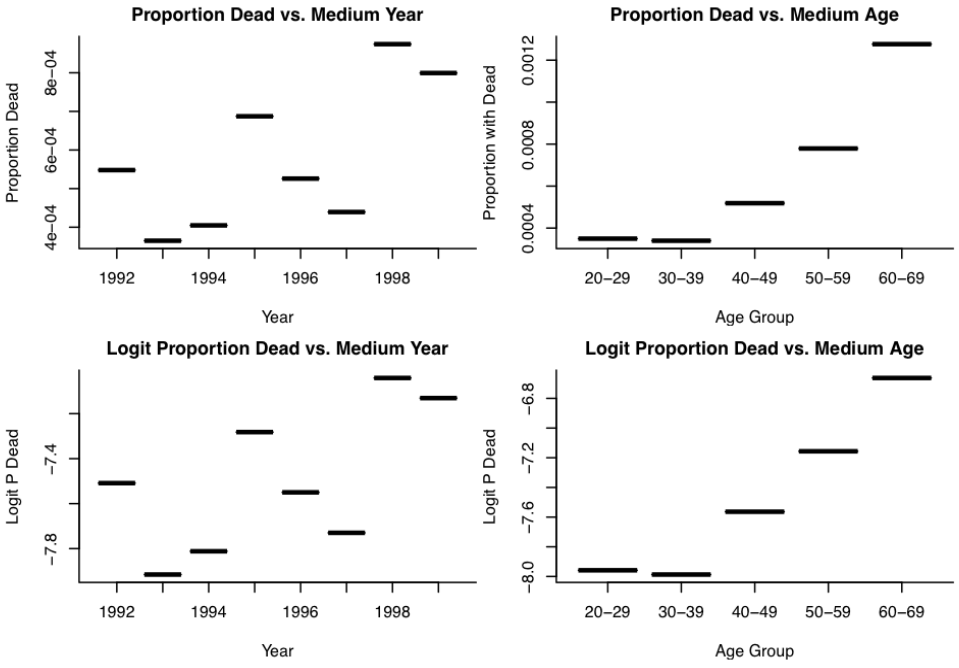
1.e

```
est.cov.beta = summary(hiro.logit.l)$cov.unscaled

a = c(0, 0 , 0 , -1, 1, 0)
ese = sqrt(t(a) %*% est.cov.beta %*% a)


lower = exp(0.704-1.96*ese)
upper = exp(0.704+1.96*ese)
```

A 95% confidence interval for the odds a patient has luekemia's ratio change when going from 100to199 to 50to99 in part d is (0.423095, 9.661593).

2.a

```
## Parsed with column specification:
## cols(
##   Year = col_integer(),
##   Numbers = col_integer(),
##   Deaths = col_integer(),
##   Age = col_character()
## )
```



2.b

```
##
## Call:
## glm(formula = prop.dead ~ as.factor(Year) + age.midpoint, family = binomial,
##     data = aviationdeaths, weights = Numbers)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.85292  -0.51150   0.09085   0.63026   2.56757
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -9.103467   0.403912 -22.538  < 2e-16 ***
## as.factor(Year)1993 -0.155832   0.331669  -0.470   0.6385
## as.factor(Year)1994 -0.096100   0.324994  -0.296   0.7675
## as.factor(Year)1995  0.379290   0.296313   1.280   0.2005
## as.factor(Year)1996  0.162257   0.315241   0.515   0.6068
## as.factor(Year)1997 -0.060501   0.350903  -0.172   0.8631
## as.factor(Year)1998  0.554678   0.298536   1.858   0.0632 .
## as.factor(Year)1999  0.436029   0.335031   1.301   0.1931
## age.midpoint         0.033084   0.006903   4.793 1.65e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

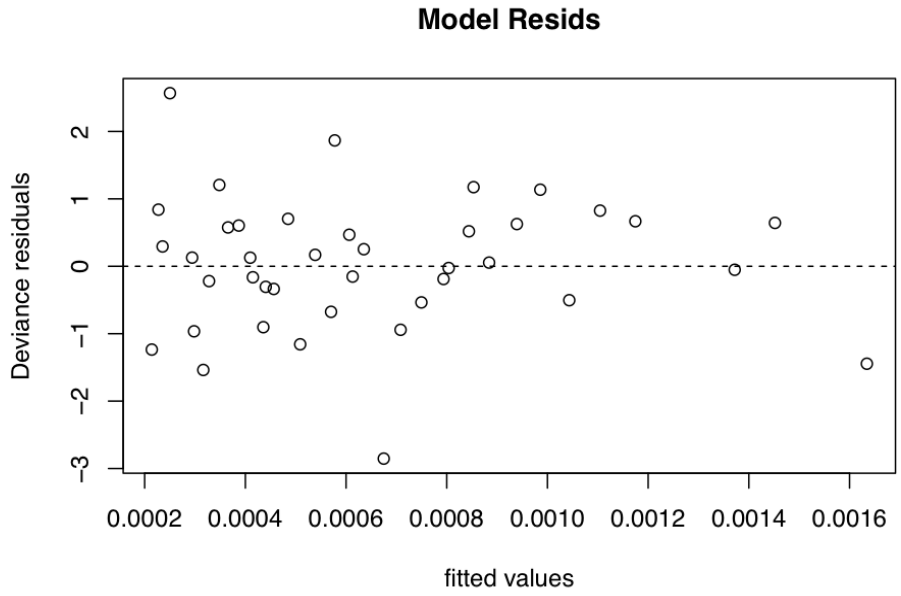
```
##     Null deviance: 74.418  on 39  degrees of freedom
## Residual deviance: 38.169  on 31  degrees of freedom
## AIC: 179.01
##
## Number of Fisher Scoring iterations: 5
```

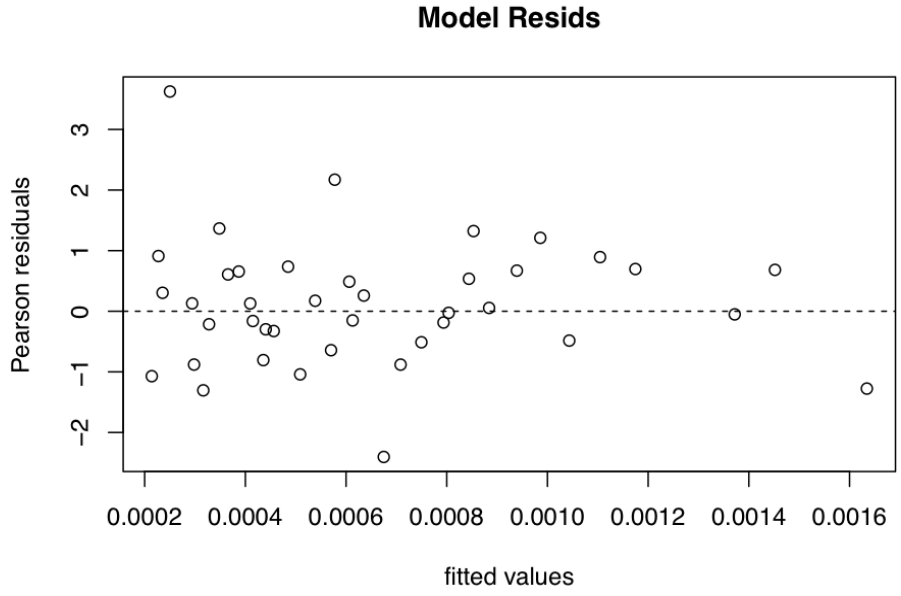
2.c

The statistical model can be written as logit(p) = B0 + B1*Midpoint + a2 + a3 + a4 + a5+ a6 + a7 + a8. B0 represents the estimated logit(p) for Age group age.midpoint of 0 for 1992 Years. a2 through a6 represent the estimated differences, given any age midpoint, in the logit(p) between pilots in year 1992 and their corresponding Year groups. a2 corresponds to Year 1993, a3 to 1994, a4 to 1995, a5 to 1996, a6 to 1997, a7 to 1998, and a8 to 1999. The effect age.midpoint has on logit(p) is independent of which year it is. In other words, for any given year group, the affect age midpoint has on the logit(p) is consistent (same slope).

2.d

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: prop.dead
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                39     74.418
## as.factor(Year)  7   12.851        32     61.567   0.07582 .
## age.midpoint     1   23.398        31     38.169 1.317e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```





Based on the coefficient summary table, only B0 and B1 are statistically different from zero. This gives us evidence to believe that their is no difference between the logit of the proportion of deaths between pilots in year group 1992 and any other pilot year group. The intercept also has a very low p-val, indicating that the estimated logit of the proportion at age midpoint zero for year 1992 is significantly different than zero (doesn't make much sense in this context). Because B1 is significantly different from zero, there is evidence that

midpoint age plays a significant role in determing logit of the proportion of pilot deaths. Though, because a2 - a8 is not significantly different from zero, we can assume that, given any midpoint age, there is no difference between the logit of the proportion of pilot deaths in year 1992 and any other year group.

The pearson and the deviance residuals look proportionately similar. Both residual plots' points are centered around 0, which is what we want to see in a good model. Another good attribute is the fact these points seem to normally distributed about zero, with the number of points increasing as you approach zero and less points further out. The fact that the density of points deacreases the larger the fitted values (x-axis) gets is not, however, preferred.

The analysis of deviance table tells us whether or not adding the given parameter contributes statistically significant data for explaining the logit of the proportion of deaths of private pilots. According to the table, adding the model including Year as a factor is not preferred over the NULL model. This is because the p-value for the f-test with the NULL model under H0 and the model including Year under H1 is larger than 0.05 (.07582). Based on the second iteration on the table, adding the model including midpoint age as a numeric is preferred over the simple logistic regression model including year as a factor. This is because the p-value for the f-test with the year as a factor only model under H0 and the model including Year and Age Group under H1 is low (1.317e-06).

2.e

```
##
## Call:
## glm(formula = PropDead ~ as.factor(Year) * age.midpoint, family = binomial,
##     data = aviationdeaths, weights = Numbers)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.57511  -0.39845  -0.01038   0.48070   2.11610
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -6.11930    0.86994  -7.034    2e-12 ***
## as.factor(Year)1993              -4.76938    1.34036  -3.558 0.000373 ***
## as.factor(Year)1994              -3.60567    1.28416  -2.808 0.004988 **
## as.factor(Year)1995              -3.03481    1.20810  -2.512 0.012003 *
## as.factor(Year)1996              -3.09592    1.23170  -2.514 0.011952 *
## as.factor(Year)1997              -3.08926    1.38221  -2.235 0.025417 *
## as.factor(Year)1998              -2.11217    1.19640  -1.765 0.077491 .
## as.factor(Year)1999              -2.43190    1.39873  -1.739 0.082096 .
## age.midpoint                     -0.03094    0.01955  -1.583 0.113502
## as.factor(Year)1993:age.midpoint  0.09939    0.02813   3.533 0.000411 ***
## as.factor(Year)1994:age.midpoint  0.07560    0.02778   2.722 0.006496 **
## as.factor(Year)1995:age.midpoint  0.07322    0.02601   2.816 0.004868 **
## as.factor(Year)1996:age.midpoint  0.07010    0.02689   2.607 0.009143 **
## as.factor(Year)1997:age.midpoint  0.06500    0.02984   2.178 0.029383 *
## as.factor(Year)1998:age.midpoint  0.05736    0.02582   2.222 0.026308 *
## as.factor(Year)1999:age.midpoint  0.06164    0.02939   2.097 0.035991 *
## ---
## 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: 24.302  on 24  degrees of freedom
## AIC: 179.14
```

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

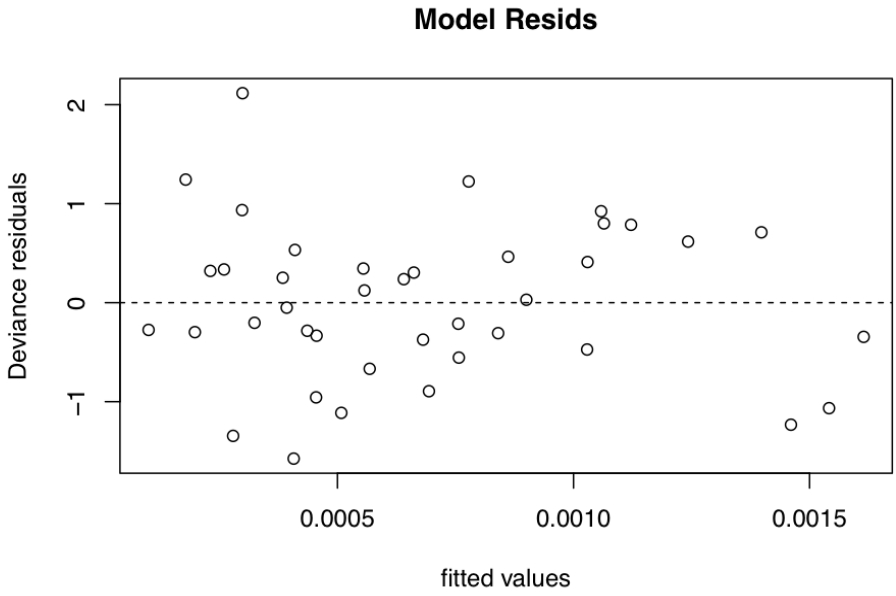
i. The statistical model can be represented as logit(p) = a + yi + (B + (By)i)age.midpointij where a + yi deals with the intercept. a = -6.1193 and yi corresponds to the shift in intercept for whichever year it represents. y1 = 0 corresponds to year group 1992. (B + (By)i)age.midpoint deals with the slope. B = -0.03094 and represents the slope for year group 1992. (By)i represents the difference in slope between year group 1992 and whichever year group it represents.

ii. The logistic regression model is not the NULL model because the number of parameters in the model does not equal the number of datapoints. We have 40 data points and 10 parameters to estimate.

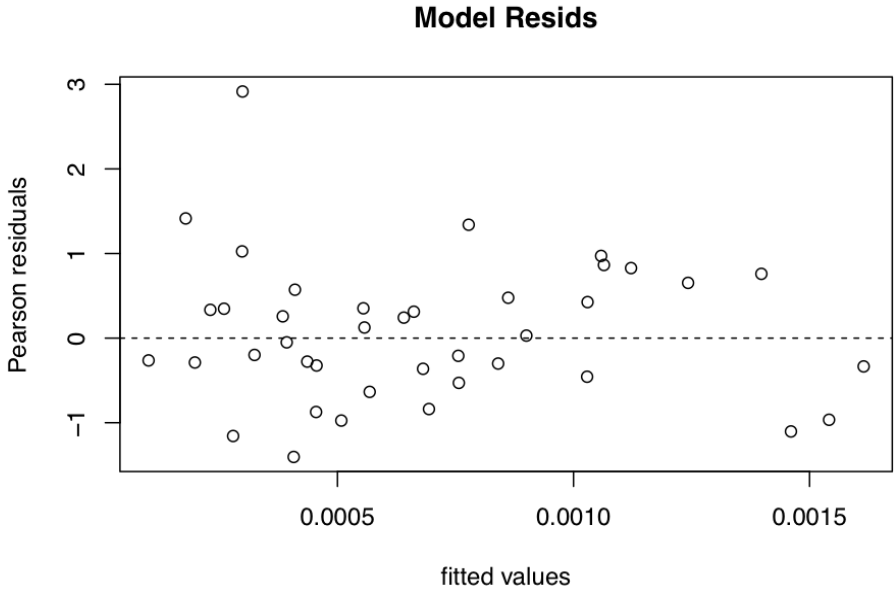
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: PropDead
##
## Terms added sequentially (first to last)
##
##
##                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                             39     74.418
## as.factor(Year)               7   12.851        32     61.567   0.07582
## age.midpoint                  1   23.398        31     38.169 1.317e-06
## as.factor(Year):age.midpoint  7   13.867        24     24.302   0.05360
##
## NULL
## as.factor(Year)              .
## age.midpoint                 ***
## as.factor(Year):age.midpoint .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

iii. According to the coefficient summary table, every parameter we estimate is significantly different from zero other than the factor coefficients representing year group 1999 and 1998 as well as the coefficient representing the B in the model, or in other words the effect age.midpoint has on the logit(p) of pilots in year group 1992. In other words, given any age.midpoint, the logit of the proportion of pilot deaths in year groups 1998 and 1999 aren't statistcally different than pilots in the 1992 year group. It also gives evidence that, for year group 1992, the midpoint of the age does not have a significant impact on the logit of the proportion of pilot deaths.

The analysis of deviance table tells us whether or not adding the given parameter contributes statistically significant data for explaining the logit of the proportion of deaths of private pilots. The first iteration gives evidence that a logistic model including Year as a factor is not more effective than just the NULL model (p val = 0.07582). But the second iteration does give evidence that the model including both Year as a factor and midpoint age as a numeric preferred over the logistic model with just the Year as a factor (p val = 1.317e-06). Finally, it gives evidence that the model including interaction effects between Year as a factor and midpoint age as a numeric is preferred over the logistic model including just Year as a factor and midpoint age as a numeric.



iv.



The data points in the residual plots appear centered about the y=0 line, which is what we want in a good predictive model. The spread is not exactly even. There seem to be more points on the left side of the x axis, particularly about 0.0005 value, which is not ideal. However, the residual points do appear to be normally distributed about zero, with more points closer to zero and less points further from zero. Given the fact that the spread of these residuals are more consistent on the fitted values(x) axis than the residuals of the model

without the interaction effects, I would preffer this model.