ST346 Coursework 2

u1801116

1(a)

Load data courseworkData2.rda.

```
load("courseworkData2.rda")
```

Fit a null and max Poisson regression model with total number of damage incidents as the outcome. Then, based on AIC, find the best model and look at the model.

```
glm.out1 <- glm(incidents ~ 1 + offset(log(service)), family=poisson(),
                data=ships)
glm.out2 <- glm(incidents ~ (type + year + period)^2 + offset(log(service)),
                family=poisson(), data=ships)
step(glm.out1, direction="both",
     scope=list("lower"=glm.out1, "upper"=glm.out2))
glm.out3 <- glm(incidents ~ year + type + period + year:type +
                offset(log(service)), family=poisson(),
                data=ships)
summary(glm.out3)
```

However, this model is clearly inappropriate. Based on AIC, we look at the next best model.

```
glm.out4 <- glm(incidents ~ year + type + period + offset(log(service)),
                family=poisson(), data=ships)
summary(glm.out4)
```

```
##
## Call:
## glm(formula = incidents ~ year + type + period + offset(log(service)),
##     family = poisson(), data = ships)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6768  -0.8293  -0.4370   0.5058   2.7912
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.40590    0.21744 -29.460  < 2e-16 ***
## year1965-69    0.69714    0.14964   4.659 3.18e-06 ***
## year1970-74    0.81843    0.16977   4.821 1.43e-06 ***
## year1975-79    0.45343    0.23317   1.945  0.05182 .
## typeB         -0.54334    0.17759  -3.060  0.00222 **
## typeC         -0.68740    0.32904  -2.089  0.03670 *
## typeD         -0.07596    0.29058  -0.261  0.79377
## typeE          0.32558    0.23588   1.380  0.16750
```

```
## period1975-79  0.38447    0.11827   3.251  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  38.695  on 25  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 154.56
##
## Number of Fisher Scoring iterations: 5
```

Hence, glm.out4 is our final model.

1(b)

Look at the summary of the model, we can see that a ship of type E, constructed on year 1970 to year 1974 and have period of operation between 1975 to 1979 have the highest risk of damage.

1(c)

Extracting coefficients from the summary of the model glm.out4.

```
expected = exp(summary(glm.out4)$coefficients["(Intercept)","Estimate"] +
               summary(glm.out4)$coefficients["typeE","Estimate"] +
               summary(glm.out4)$coefficients["year1970-74","Estimate"] +
               summary(glm.out4)$coefficients["period1975-79","Estimate"])*
  3353*(1/5)
expected
```

```
## [1] 5.107675
```

The expected number of damage incidents in twelve months for the class of shipping at 1(b) is `5.1076748`.

2(a)

Exponential Dispersion Model: Binomial

Link function: Logit

Outcome variable: Number of school leavers who have a place at university or other higher education institute divided by total number of school leavers

Predictor variable: age, gender and smoking status

2(b)

Exponential Dispersion Model: Normal

Link function: Identity

Outcome variable: change in weights of individual rats after six month separated in two groups

Predictor variable: vitamin D intake in factor

2(c)

Exponential Dispersion Model: Poisson

Link function: Log

Outcome variable: log of number of times of break downs divided by number of miles driven

Predictor variable: model of the vehicle and number of times that it has been serviced

3(a)

Add a new column in the data frame that represents the total number of participants for each group.

```
esoph$participants <- esoph$ncases + esoph$ncontrols
```

Fit a null and max Binomial regression model with percentage of participants with esophageal cancer as the outcome. Then, based on AIC, find the best model.

```
glm.out5 <- glm(ncases/participants ~ 1, family=binomial(),
                weight=participants,data=esoph)
glm.out6 <- glm(ncases/participants ~ (agegp + alcgp + tobgp)^2,
                family=binomial(), weight=participants, data=esoph)
step(glm.out5, direction="both",
     scope=list("lower"=glm.out5, "upper"=glm.out6))
```

We now look at the model.

```
glm.out7 <- glm(ncases/participants ~ alcgp + agegp + tobgp,
                family=binomial(), weight=participants,
                data=esoph)
summary(glm.out7)
```

```
##
## Call:
## glm(formula = ncases/participants ~ alcgp + agegp + tobgp, family = binomial(),
##     data = esoph, weights = participants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9507  -0.7376  -0.2438   0.6130   2.4127
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.8954     1.0859  -6.350 2.16e-10 ***
## alcgp40-79    1.4346     0.2501   5.737 9.63e-09 ***
## alcgp80-119   1.9807     0.2848   6.956 3.51e-12 ***
## alcgp120+     3.6029     0.3850   9.357  < 2e-16 ***
## agegp35-44    1.9809     1.1041   1.794 0.072786 .
## agegp45-54    3.7763     1.0680   3.536 0.000407 ***
## agegp55-64    4.3352     1.0650   4.070 4.69e-05 ***
```

```
## agegp65-74    4.8964     1.0764   4.549 5.39e-06 ***
## agegp75+      4.8265     1.1213   4.304 1.67e-05 ***
## tobgp10-19    0.4381     0.2283   1.919 0.055039 .
## tobgp20-29    0.5126     0.2730   1.878 0.060398 .
## tobgp30+      1.6410     0.3441   4.769 1.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  82.337  on 76  degrees of freedom
## AIC: 221.39
##
## Number of Fisher Scoring iterations: 6
```

Hence, glm.out7 is our best model.

3(b)

```
glm.out8 <- glm(ncases/participants ~ as.numeric(alcgp) + agegp + tobgp,
                family=binomial(), weight=participants, data=esoph)
anova(glm.out8, glm.out5, "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: ncases/participants ~ as.numeric(alcgp) + agegp + tobgp
## Model 2: ncases/participants ~ 1
##   Resid. Df Resid. Dev Df Deviance
## 1        78      88.24
## 2        87     367.95 -9  -279.71
```

```
glm.out9 <- glm(ncases/participants ~ alcgp + as.numeric(agegp) + tobgp,
                family=binomial(), weight=participants, data=esoph)
anova(glm.out9, glm.out5, "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: ncases/participants ~ alcgp + as.numeric(agegp) + tobgp
## Model 2: ncases/participants ~ 1
##   Resid. Df Resid. Dev Df Deviance
## 1        80     101.89
## 2        87     367.95 -7  -266.06
```

```
glm.out10 <- glm(ncases/participants ~ alcgp + agegp + as.numeric(tobgp),
                 family=binomial(), weight=participants, data=esoph)
anova(glm.out10, glm.out5, "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: ncases/participants ~ alcgp + agegp + as.numeric(tobgp)
## Model 2: ncases/participants ~ 1
##   Resid. Df Resid. Dev Df Deviance
## 1        78      85.87
```

```
## 2        87     367.95 -9  -282.09
```

```
summary(glm.out10)
```

```
##
## Call:
## glm(formula = ncases/participants ~ alcgp + agegp + as.numeric(tobgp),
##     family = binomial(), data = esoph, weights = participants)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2009  -0.7033  -0.1997   0.4880   2.3644
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -7.1682     1.0910  -6.570 5.02e-11 ***
## alcgp40-79          1.3941     0.2479   5.623 1.88e-08 ***
## alcgp80-119         1.9701     0.2817   6.993 2.69e-12 ***
## alcgp120+           3.5715     0.3802   9.393  < 2e-16 ***
## agegp35-44          1.8180     1.0931   1.663 0.096272 .
## agegp45-54          3.6212     1.0545   3.434 0.000595 ***
## agegp55-64          4.1696     1.0506   3.969 7.23e-05 ***
## agegp65-74          4.7168     1.0610   4.445 8.77e-06 ***
## agegp75+            4.6941     1.1089   4.233 2.31e-05 ***
## as.numeric(tobgp)   0.4322     0.0967   4.470 7.83e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 367.953  on 87  degrees of freedom
## Residual deviance:  85.866  on 78  degrees of freedom
## AIC: 220.92
##
## Number of Fisher Scoring iterations: 6
```

From the ANOVA tables, we can see that the p-value for `glm.out10` is 0.17, this means that there is no evidence to reject `glm.out10` from the null model. The variable `tobgp` should be a numeric value.

3(c)

```
odds_ratio = exp(summary(glm.out10)$coefficients["alcgp120+","Estimate"])
odds_ratio
```

```
## [1] 35.57026
```

The odds ratio of esophageal cancer for a participant who drinks more than 120g of alcohol/day compared with one who drinks less than 40g of alcohol per day is `35.5702575`.

4