ST346: Assessed coursework 1

u1801116

1(a)

Load data courseworkData1.rda, then fit a null Poisson regression model with number of claims as the outcome and name it `glm.out1`.

```
load("courseworkData1.rda")
glm.out1 = glm(formula = y ~ 1 + offset(log(n)), family = "poisson", data = insurance)
glm.out1
```

```
##
## Call:  glm(formula = y ~ 1 + offset(log(n)), family = "poisson", data = insurance)
##
## Coefficients:
## (Intercept)
##      -2.003
##
## Degrees of Freedom: 31 Total (i.e. Null);  31 Residual
## Null Deviance:       207.8
## Residual Deviance: 207.8     AIC: 378.2
```

As we can see from the summary of `glm.out1`, the intercept of the model is -2.00326, now check it with equation.

```
intercept = log(sum(insurance$y)/sum(insurance$n))
intercept
```

```
## [1] -2.003262
```

The equation also equals to -2.00326.

1(b)

Fit a Poisson regression model with predictor variables `car`, `age`, `district`.

```
glm.out2 = glm(formula = y ~ factor(car) + factor(age) + district + offset(log(n)), family = "poisson",
modelsum <- summary(glm.out2)
exp(modelsum$coefficients["district","Estimate"])
```

```
## [1] 1.244203
```

An estimate of the rate ratio is `1.2442031`, the rate of insurance claims is higher in urban areas than rural areas.

1(c)

```r
glm.max = glm(formula = y ~ (factor(car) + factor(age) + district + offset(log(n)))^2, family = "poisson
step(glm.max, direction = "both")
```

As we shown from the results, AIC is minimal at 208.07 when model is the same as the one in 1(b), therefore the maximal model is the one shown in 1(b).

1(d)

```r
glm.out3 = glm(formula = y ~ factor(car) + age + district + offset(log(n)), family = "poisson", data = i
anovatable = anova(glm.out2, glm.out3, test = "LRT")
anovatable
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ factor(car) + factor(age) + district + offset(log(n))
## Model 2: y ~ factor(car) + age + district + offset(log(n))
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        24     23.709
## 2        26     23.832 -2  -0.1234   0.9402
```

As Pr(>Chi) is `0.9401661`, according to 0.05 level of significance, there are no significant differences between the two model.

1(e)

The used model is `glm.out3`.

```r
summary(glm.out3)
```

```
##
## Call:
## glm(formula = y ~ factor(car) + age + district + offset(log(n)),
##     family = "poisson", data = insurance)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8383  -0.5899  -0.1651   0.3733   1.7783
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.63733    0.07499 -21.833  < 2e-16 ***
## factor(car)2  0.16260    0.05048   3.221 0.001276 **
## factor(car)3  0.39389    0.05491   7.174 7.31e-13 ***
## factor(car)4  0.56585    0.07216   7.842 4.44e-15 ***
## age          -0.17616    0.01850  -9.523  < 2e-16 ***
## district      0.21860    0.05853   3.735 0.000188 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance: 207.833  on 31  degrees of freedom
## Residual deviance:  23.832  on 26  degrees of freedom
## AIC: 204.19
##
## Number of Fisher Scoring iterations: 4
```

```
exp(summary(glm.out3)$coefficients["factor(car)4","Estimate"])
```

```
## [1] 1.760939
```

Therefore, the new insurance premium is 1.7609386.

2(a)

```
library(dplyr)
glm.out4 = glm(deaths ~ smoking + offset(log(personyears)),
               family = "poisson", data = doctors)
summary(glm.out4)
```

```
##
## Call:
## glm(formula = deaths ~ smoking + offset(log(personyears)), family = "poisson",
##     data = doctors)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -16.535   -6.031    4.612    8.162   13.644
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.9618     0.0995 -59.916  < 2e-16 ***
## smoking       0.5422     0.1072   5.059 4.22e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 935.07  on 9  degrees of freedom
## Residual deviance: 905.98  on 8  degrees of freedom
## AIC: 965.04
##
## Number of Fisher Scoring iterations: 6
```

```
smoker = filter(doctors, smoking == 1)
nosmoker = filter(doctors, smoking == 0)
lambda1 = sum(smoker$deaths)/sum(smoker$personyears)
lambda0 = sum(nosmoker$deaths)/sum(nosmoker$personyears)
log(lambda1/lambda0)
```

```
## [1] 0.5422211
```

As we can see, the value is the same as the estimate of coefficient for smoking.

2(b)

```r
glm.out5 = glm(deaths ~ factor(age) + smoking + offset(log(personyears)),
               family = "poisson", data = doctors)
summary(glm.out5)$coefficients["smoking","Estimate"]
```

```
## [1] 0.3545356
```

The estimate of β drops to `0.3545356`.

2(c)

Shown from the two diagrams, the model in 2b is clearly not appropriate as there is a relationship between age and smoking. Using stratified parameterization for the new model.

```r
glm.out6 = glm(deaths ~ (factor(age)/factor(smoking)) + offset(log(personyears)),
               family = "poisson", data = doctors)
summary(glm.out6)
```

```
##
## Call:
## glm(formula = deaths ~ (factor(age)/factor(smoking)) + offset(log(personyears)),
##     family = "poisson", data = doctors)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          -9.1479     0.7071 -12.937  < 2e-16 ***
## factor(age)45 to 54                   2.3574     0.7638   3.087  0.00203 **
## factor(age)55 to 64                   3.8302     0.7319   5.233 1.67e-07 ***
## factor(age)65 to 74                   4.6227     0.7319   6.316 2.69e-10 ***
## factor(age)75 to 84                   5.2944     0.7296   7.257 3.96e-13 ***
## factor(age)35 to 44:factor(smoking)1  1.7469     0.7289   2.397  0.01654 *
## factor(age)45 to 54:factor(smoking)1  0.7603     0.3049   2.494  0.01264 *
## factor(age)55 to 64:factor(smoking)1  0.3841     0.2014   1.907  0.05654 .
## factor(age)65 to 74:factor(smoking)1  0.3046     0.2027   1.503  0.13295
## factor(age)75 to 84:factor(smoking)1 -0.1001     0.2051  -0.488  0.62543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 9.3507e+02  on 9  degrees of freedom
## Residual deviance: 4.4409e-16  on 0  degrees of freedom
## AIC: 75.068
##
## Number of Fisher Scoring iterations: 3
```

```r
exp(summary(glm.out6)$coefficients["factor(age)65 to 74:factor(smoking)1","Estimate"])
```

```
## [1] 1.35606
```

The mortality rate ratio is `1.3560598` and p-value is `0.13295`.