ActSc 632 Test 2 Solutions

Spring 2023

Department of Statistics and Actuarial Science, University of Waterloo

```
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
```

Question 1

By examining the *KenyaCarInsurance* data set, state how many tariff cells there are. Also, determine how many rating factors we have, as well as the number of categories for each rating factor.

Solution

```
KenyaCarInsurance<-read.csv("KenyaCarInsurance.csv",header=TRUE,sep=',')

head(KenyaCarInsurance, 3)
```

```
##   Age_group Gender License_Type                      Category
## 1         1      1            1         17 to 20, Female, Full
## 2         1      1            2 17 to 20, Female, Provisional
## 3         1      2            1           17 to 20, Male, Full
##   Earned.Premium.Income..E.000. exposure_years total_cost nb_claims  severity
## 1                       4312462        4426.37    2816708       352  8002.011
## 2                       4994867        3482.29    4123536       413  9984.349
## 3                       4826916        3403.84    3652734       269 13578.937
```

```
KenyaCarInsurance <- within(KenyaCarInsurance, {

    age <- factor(Age_group)

    gender <- factor(Gender)

    license <- factor(License_Type)

})

nrow(KenyaCarInsurance)
```

```
## [1] 24
```

The number of tariff cells is 24. There are 3 rating factors age group, gender and license_type, and the number of categories for each rating factor are

```
levels(KenyaCarInsurance$age)
```

```
## [1] "1" "2" "3" "4" "5" "6"
```

```
levels(KenyaCarInsurance $gender)
```

```
## [1] "1" "2"
```

```
levels(KenyaCarInsurance $license)
```

```
## [1] "1" "2"
```

The predictor age has 6 categories, while the predictors gender and license-type have 2 categories

For each tariff cell, we have information about duration (the number of policyholder years of experience within the tariff cell), number (number of claims in the tariff cell) and severity (average claim for each tariff cell)

Question 2

You are asked to set as the base tariff cell (for both the frequency and severity models) the one with the largest duration. Identify explicitly this base tariff cell.

Solution

The new basecell is

```
print(basecell <- KenyaCarInsurance[which.max(KenyaCarInsurance$exposure_years), 1:3])
```

```
##    Age_group Gender License_Type
## 13         4      1            1
```

```
KenyaCarInsurance$age<-relevel(KenyaCarInsurance$age, as.character(basecell$Age_group))

KenyaCarInsurance$gender<-relevel(KenyaCarInsurance$gender, as.character(basecell$Gender))

KenyaCarInsurance$license<-relevel(KenyaCarInsurance$license, as.character(basecell$License_Type))
```

Question 3

Model the frequency data using a glm based on the relative Poisson distribution and a log-link function. Using the concept of deviance, discuss the overall fit of this model. Does the relative Poisson glm model seem reasonable to fit the frequency data? Comment.

Solution

```
summary(freq <-glm(nb_claims ~ age + gender + license + offset(log(exposure_years)), family=poisson("log"), data=
KenyaCarInsurance[KenyaCarInsurance$exposure_years > 0,]))
```

```
##
## Call:
## glm(formula = nb_claims ~ age + gender + license + offset(log(exposure_years)),
##     family = poisson("log"), data = KenyaCarInsurance[KenyaCarInsurance$exposure_years >
##         0, ])
##
## Coefficients:
##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.820168   0.006097 -462.518  < 2e-16 ***
## age1         0.308509   0.030273   10.191  < 2e-16 ***
## age2         0.170420   0.016575   10.282  < 2e-16 ***
## age3         0.110777   0.010922   10.142  < 2e-16 ***
## age5        -0.228286   0.008548  -26.705  < 2e-16 ***
## age6        -0.168754   0.014977  -11.268  < 2e-16 ***
## gender2      0.023309   0.007117    3.275  0.00106 **
## license2     0.340993   0.013077   26.075  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2431.814  on 23  degrees of freedom
## Residual deviance:   56.033  on 16  degrees of freedom
## AIC: 282.35
##
## Number of Fisher Scoring iterations: 3
```

The scaled deviance is

```
cbind(scaled.deviance = freq$deviance,
      df = freq$df.residual,
      p = 1 - pchisq(freq$deviance, freq$df.residual))
```

```
##      scaled.deviance df            p
## [1,]        56.03298 16 2.403233e-06
```

The fit of the model as measured by the deviance is not good. The (scaled) deviance statistic is

```
freq$deviance
```

```
## [1] 56.03298
```

which is above the critical value at a 95% confidence level of a chi-square rv with

```
freq$df.residual
```

```
## [1] 16
```

degrees of freedom, which is

```
qchisq(0.95, freq$df.residual)
```

```
## [1] 26.29623
```

As such, we have evidence to reject the null hypothesis and we therefore conclude that the relative Poisson model does not provide a good fit for the data set in question.

Question 4

Now model the frequency data using a glm based on the relative quasi-Poisson distribution, still with a log-link function. Using the concept of deviance, discuss the overall fit of this model. Does the relative quasi-Poisson glm model seem reasonable to fit the frequency data? Comment.

Solution

Below is a trained quasi-Poisson glm model

```
summary(freqqp <-glm(nb_claims ~ age + gender + license + offset(log(exposure_years)), family=quasipoisson("lo
g"), data=KenyaCarInsurance[KenyaCarInsurance$exposure_years > 0,]))
```

```
##
## Call:
## glm(formula = nb_claims ~ age + gender + license + offset(log(exposure_years)),
##     family = quasipoisson("log"), data = KenyaCarInsurance[KenyaCarInsurance$exposure_years >
##         0, ])
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.82017    0.01129 -249.821  < 2e-16 ***
## age1         0.30851    0.05605    5.504 4.80e-05 ***
## age2         0.17042    0.03069    5.554 4.36e-05 ***
## age3         0.11078    0.02022    5.478 5.06e-05 ***
## age5        -0.22829    0.01583  -14.424 1.37e-10 ***
## age6        -0.16875    0.02773   -6.086 1.58e-05 ***
## gender2      0.02331    0.01318    1.769   0.0959 .
## license2     0.34099    0.02421   14.084 1.96e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.427685)
##
##     Null deviance: 2431.814  on 23  degrees of freedom
## Residual deviance:   56.033  on 16  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 3
```

with estimated dispersion

```
print(freqqp.phi<-summary(freqqp)$dispersion)
```

```
## [1] 3.427685
```

we compute the scaled deviance and this time we do not reject the null hypothesis that the quais-Poisson model provides a good fit for the data, as our p-value is now 0.429.

```
cbind(scaled.deviance=freqqp$deviance/freqqp.phi,df=freqqp$df.residual,p=1-pchisq(freqqp$deviance/freqqp.phi,freq
qp$df.residual))
```

```
##      scaled.deviance df         p
## [1,]        16.34718 16 0.4290038
```

Question 5

Propose a potential simplification to the relative quasi-Poisson glm model of Q4 by dropping one of the rating factors and explain your choice. Via a likelihood ratio test argument, determine whether or not you are statistically justified to simplify the relative quasi-Poisson glm model of Q4. State the null and alternative hypothesis of this likelihood ratio test together with its test statistics. Continue the test with the model chosen in this sub-question.

Solution

Suggestion to simplify the relative Poisson glm model: based on the parameter estimation, we suggest to remove gender.

```
summary(freqqp1 <-glm(nb_claims~ age + license + offset(log(exposure_years)), family=quasipoisson("log"), data=Ke
nyaCarInsurance[KenyaCarInsurance$exposure_years > 0,]))
```

```
##
## Call:
## glm(formula = nb_claims ~ age + license + offset(log(exposure_years)),
##     family = quasipoisson("log"), data = KenyaCarInsurance[KenyaCarInsurance$exposure_years >
##         0, ])
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.80965    0.01015 -276.815  < 2e-16 ***
## age1         0.30720    0.05943    5.169 7.71e-05 ***
## age2         0.16877    0.03253    5.189 7.40e-05 ***
## age3         0.10964    0.02143    5.116 8.61e-05 ***
## age5        -0.22563    0.01671  -13.505 1.62e-10 ***
## age6        -0.16491    0.02931   -5.626 3.03e-05 ***
## license2     0.33817    0.02562   13.200 2.31e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 3.854584)
##
##     Null deviance: 2431.814  on 23  degrees of freedom
## Residual deviance:   66.756  on 17  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 3
```

The likelihood ratio is:

- H_0: beta of the rating factor “gender” is 0
- H_a: this betas is different than 0

```
print(ltest1<-anova(freqqp1,freqqp))
```

```
## Analysis of Deviance Table
##
## Model 1: nb_claims ~ age + license + offset(log(exposure_years))
## Model 2: nb_claims ~ age + gender + license + offset(log(exposure_years))
##   Resid. Df Resid. Dev Df Deviance
## 1        17     66.756
## 2        16     56.033  1   10.723
```

```
qchisq(0.95,ltest1$Df[2])
```

```
## [1] 3.841459
```

with the scaled deviance of

```
ltest1$Deviance[2]/freqqp.phi
```

```
## [1] 3.128277
```

```
1-pchisq(ltest1$Deviance[2]/freqqp.phi,ltest1$Df[2])
```

```
## [1] 0.07694504
```

The (scaled) deviance moves up from

```
freqqp$deviance/freqqp.phi
```

```
## [1] 16.34718
```

for the model in Q4 to

```
freqqp1$deviance/freqqp.phi
```

```
## [1] 19.47546
```

When the predictor “gender” is not included in the relative Poisson glm model. Given that

```
ltest1$Df[2]
```

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

degrees of freedom are gained with the simplified model, we shall compare the gain in deviance of

```
freqqp1$deviance/freqqp.phi - freqqp$deviance/freqqp.phi
```

```
## [1] 3.128277
```

to the critical value at a 95% confidence level of a chi-square rv with

```
ltest1$Df[2]
```

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

degree of freedom, which is

```
qchisq(0.95,ltest1$Df[2])
```

```
## [1] 3.841459
```

As such, we are statistically justified to simplify the modelby excluding the predictor “gender” from the glm model.

Question 6

Using the relative quasi-Poisson glm model chosen in Q5:

a. estimate the multipliers (i.e., relativities) for the frequency data and provide a 95% confidence interval for each.

b. identify the tariff cell with the largest expected number of claims per exposure period. Find the expected number of claims per year for a policy in this tariff cell.

Solution

Part (a)

Relativities and CI (at a 95% confidence level) for the relative Poisson glm model is

```
cbind(exp(freqqp1$coefficients),exp(freqqp1$coefficients-qnorm(0.975)*sqrt(diag(vcov(freqqp1)))),exp(freqqp1$coef
ficients+qnorm(0.975)*sqrt(diag(vcov(freqqp1)))))
```

```
##                   [,1]       [,2]       [,3]
## (Intercept) 0.06022624 0.05903997 0.06143634
## age1        1.35961176 1.21011072 1.52758265
## age2        1.18385061 1.11073470 1.26177948
## age3        1.11588103 1.06997631 1.16375518
## age5        0.79801393 0.77230566 0.82457796
## age6        0.84797160 0.80062525 0.89811787
## license2    1.40237397 1.33369551 1.47458903
```

Part (b)

The tariff cell with the most expected claim per exposure period is age group 1 and license 2. You identify this tariff cell by picking for each rating factor the beta which is the greatest The expected number of claim over one policy yearsfor a policyholder in this tariff cell is

```
prod(exp(freqqp1$coefficients)[c(1,2,7)])
```

```
## [1] 0.1148324
```

The tariff cell with the less expected claim per exposure period is age group 5 and license 1. You identify this tariff cell by picking for each rating factor the beta which is the smallest. The expected number of claim over two policy years for a policyholder in this tariff cell is

```
prod(exp(freqqp1$coefficients)[c(1,5)] )
```

```
## [1] 0.04806138
```

Question 7

Model the severity data using a glm based on the gamma distribution and a log-link function. Using the concept of deviance, discuss the overall fit of the model. Does the gamma glm model seem reasonable to model the claim severity data set? Comment.

Solution

Below is a trained gamma glm model:

```
summary(sev <- glm(severity ~ age + gender + license, family = Gamma("log"), data = KenyaCarInsurance[KenyaCarIns
urance$nb_claims > 0, ], weights = nb_claims))
```

```
##
## Call:
## glm(formula = severity ~ age + gender + license, family = Gamma("log"),
##     data = KenyaCarInsurance[KenyaCarInsurance$nb_claims > 0,
##         ], weights = nb_claims)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.48876    0.03586 236.720  < 2e-16 ***
## age1         0.57819    0.17844   3.240  0.00513 **
## age2         0.34541    0.09768   3.536  0.00275 **
## age3         0.12428    0.06431   1.933  0.07119 .
## age5        -0.15699    0.05032  -3.120  0.00660 **
## age6         0.02776    0.08813   0.315  0.75683
## gender2      0.15900    0.04182   3.802  0.00157 **
## license2     0.25048    0.07727   3.241  0.00511 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 34.64568)
##
##     Null deviance: 3259.00  on 23  degrees of freedom
## Residual deviance:  544.44  on 16  degrees of freedom
## AIC: 1220444
##
## Number of Fisher Scoring iterations: 5
```

```
sev.phi<-summary(sev)$dispersion

cbind(res.deviance = sev$deviance/sev.phi, df = sev$df.residual, p = 1-pchisq(sev$deviance/sev.phi, sev$df.residu
al))
```

```
##      res.deviance df        p
## [1,]     15.71464 16 0.473048
```

As measured by the deviance statistic, the gamma glm fit seems to be quite good. We obtain an unscaled deviance of

```
sev$deviance
```

```
## [1] 544.4445
```

with an estimated dispersion parameter of

```
sev.phi
```

```
## [1] 34.64568
```

which leads to a scaled deviance of

```
sev$deviance/sev.phi
```

```
## [1] 15.71464
```

This should be compared with the critical value at a 95% confidence level of a chi-square rv with

```
sev$df.residual
```

```
## [1] 16
```

degrees of freedom, which is

```
qchisq(0.95,sev$df.residual)
```

```
## [1] 26.29623
```

yielding a p-value for the test of

```
1-pchisq(sev$deviance/sev.phi, sev$df.residual)
```

```
## [1] 0.473048
```

The gamma glm seems to offer a very good fit to the data.

Question 8

You are now asked to combine the age category 6 (70 and above) with the age category 4 (31-50) in the gamma glm model of Q7. Using a likelihood ratio test, comment on the appropriateness to do so. State the null and alternative hypothesis of this likelihood ratio test. Are you statistically justified to simplify the model in Q7 to the one suggested here? Comment.

Solution

Merge categories 6 with 4 for age:

```
levels(KenyaCarInsurance$age)<-recode(levels(KenyaCarInsurance$age), "4"="4 & 6")
levels(KenyaCarInsurance$age)<-recode(levels(KenyaCarInsurance$age), "6"="4 & 6")

levels(KenyaCarInsurance$age)<-recode(levels(KenyaCarInsurance$age), "5"="5")

summary(sev1 <- glm(severity~ age + gender + license, family = Gamma("log"), data = KenyaCarInsurance[KenyaCarIns
urance$nb_claims > 0, ], weights = nb_claims))
```

```
##
## Call:
## glm(formula = severity ~ age + gender + license, family = Gamma("log"),
##     data = KenyaCarInsurance[KenyaCarInsurance$nb_claims > 0,
##         ], weights = nb_claims)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.49161    0.03385 250.849  < 2e-16 ***
## age1         0.57544    0.17393   3.308  0.00415 **
## age2         0.34238    0.09488   3.609  0.00217 **
## age3         0.12117    0.06202   1.954  0.06739 .
## age5        -0.16031    0.04800  -3.340  0.00388 **
## gender2      0.16004    0.04070   3.932  0.00107 **
## license2     0.24923    0.07532   3.309  0.00415 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 32.98963)
##
##     Null deviance: 3259.0  on 23  degrees of freedom
## Residual deviance:  547.9  on 17  degrees of freedom
## AIC: 1220958
##
## Number of Fisher Scoring iterations: 5
```

Likelihood ratio test:

```
print(ltest3<-anova(sev1,sev))
```

```
## Analysis of Deviance Table
##
## Model 1: severity ~ age + gender + license
## Model 2: severity ~ age + gender + license
##   Resid. Df Resid. Dev Df Deviance
## 1        17     547.90
## 2        16     544.44  1   3.4558
```

Compare the critical value of

```
qchisq(0.95,ltest3$Df[2])
```

```
## [1] 3.841459
```

with the scaled deviance of

```
ltest3$Deviance[2]/sev.phi
```

```
## [1] 0.09974667
```

Otherwise, the p-value of the test is

```
1-pchisq(ltest3$Deviance[2]/sev.phi,ltest3$Df[2])
```

```
## [1] 0.7521339
```

The likelihood ratio test consists of * H0: combined beta for age group 4 and age group 6
* Ha: More general model in Q7

The unscaled deviance moves up from

```
sev$deviance
```

```
## [1] 544.4445
```

for the model in Q7 to

```
sev1$deviance
```

```
## [1] 547.9002
```

for the model in Q8. Given that

```
ltest3$Df[2]
```

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

degrees of freedom are gained with the simplified model, we shall compare the gain in scaled deviance of

```
(sev1$deviance-sev$deviance)/sev.phi
```

```
## [1] 0.09974667
```

to the critical value at a 95% confidence level of a chi-square rv with

```
ltest3$Df[2]
```

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

degrees of freedom, which is

```
qchisq(0.95,ltest3$Df[2])
```

```
## [1] 3.841459
```

As such, we are statistically justified to simplify the model in Q7 to the model in Q8.

Question 9

Based on your recommendation in Q8:

a. estimate the multipliers (i.e., relativities) for the severity data and provide a 95% confidenceinterval for each.

b. identify the tariff cell with the largest expected claim size. Find this expected claim size.

c. identify the tariff cell with the smallest expected claim size. Find this expected claim size.

Multipliers with 95% confidence interval :

```
cbind(exp(sev1$coefficients),exp(sev1$coefficients-qt(0.975,sev1$df.residual)*sqrt(diag(vcov(sev1)))),exp(sev1$co
efficients+qt(0.975,sev1$df.residual)*sqrt(diag(vcov(sev1)))))
```

```
##                      [,1]         [,2]         [,3]
## (Intercept) 4873.6994361 4537.7578128 5234.5116627
## age1           1.7779074    1.2317895    2.5661484
## age2           1.4082886    1.1528129    1.7203805
## age3           1.1288136    0.9903655    1.2866161
## age5           0.8518766    0.7698357    0.9426606
## gender2        1.1735540    1.0769824    1.2787851
## license2       1.2830349    1.0945228    1.5040147
```

The tariff cell with the largest expected claim amount is age = 1, gender = 2 and license is 2

```
prod(exp(sev1$coefficients)[c(1,2,6,7)])
```

```
## [1] 13046.96
```

while the tariff cell with the less expected claim amount is age = 5, gender = 1, license = 1

```
prod(exp(sev1$coefficients)[c(1,5)] )
```

```
## [1] 4151.79
```

Question 10

For the tariff cell age = 2, gender = male (1) , license = 2 the expected key frequency ratio is

```
prod(exp(freqqp1$coefficients)[c(1,3,7)])
```

```
## [1] 0.09998768
```

while the expected severity key ratio is

```
prod(exp(sev1$coefficients)[c(1,3,6,7)])
```

```
## [1] 10334.56
```

for a pure premium of

```
prod(exp(freqqp1$coefficients)[c(1,3,7)])*prod(exp(sev1$coefficients)[c(1,3,6,7)])
```

```
## [1] 1033.329
```

For the tariff cell age = 6, gender = female and license = 1 the expected key frequency ratio is

```
prod(exp(freqqp1$coefficients)[c(1,6)])
```

```
## [1] 0.05107014
```

while the expected severity key ratio is

```
prod(exp(sev1$coefficients)[c(1)])
```

```
## [1] 4873.699
```

for a pure premium of

```
prod(exp(freqqp1$coefficients)[c(1,6)])*prod(exp(sev1$coefficients)[c(1)])
```

```
## [1] 248.9005
```

Therefore, the expected pure premium for the 60 policies is

```
20*prod(exp(freqqp1$coefficients)[c(1,3,7)])*prod(exp(sev1$coefficients)[c(1,3,6,7)])+40*prod(exp(freqqp1$coeffic
ients)[c(1,6)])*prod(exp(sev1$coefficients)[c(1)])
```

```
## [1] 30622.59
```

For the set of policies, the expected number of claims is

```
print(meannb<-20*prod(exp(freqqp1$coefficients)[c(1,3,7)])+40*prod(exp(freqqp1$coefficients)[c(1,6)]))
```

```
## [1] 4.042559
```