ActSc 632 Assignment 1 Solutions

Spring 2023

Department of Statistics and Actuarial Science, University of Waterloo

```
# Install the packages if necessary
# install.packages("insuranceData")
# install.packages("doParallel")
# install.packages("dplyr")
# install.packages("tibble")

library(insuranceData)
library(foreach)
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
```

```
library(tibble)

options(tibble.print_max = Inf)
```

Question 1

Download the data set dataOhlsson and briefly discuss the data found in there (e.g., how many rating factors, what are the levels for each, how is the exposure determined, etc.). If there are any problems with the data, explain how you dealt with them.

Solution:

As stated in the R documentation for this dataset:

> The data for this case study comes from the former Swedish insurance company Wasa, and concerns partial casco insurance, for motorcycles this time. It contains aggregated data on all insurance policies and claims during 1994-1998; the reason for using this rather old data set is confidentiality; more recent data for ongoing business can not be disclosed.

```
# Load the data
data(dataOhlsson)

# change categorical columns into factors
dataOhlsson <- within(dataOhlsson, {
    zon <- factor(zon)
    mcklass <- factor(mcklass)
    bonuskl <- factor(bonuskl)
    kon <- factor(kon)})

# regroup the column bonuskl into 3 levels
levels(dataOhlsson$bonuskl) <- list("1" = c("1","2"),
                                    "2" = c("3","4"),
                                    "3" = c("5","6","7"))

# regroup the column mcklass into 3 levels
# For details, visit: https://stackoverflow.com/questions/13559076/group-numeric-values-by-the-intervals
dataOhlsson$motorage_gr <- cut(dataOhlsson$fordald,
                               breaks = c(0, 2, 5, 100),
                               labels = c("1", "2", "3"),
                               right = FALSE)

summary(dataOhlsson)
```

```
##      agarald      kon       zon       mcklass      fordald      bonuskl
##  Min.   : 0.00   K: 9853   1: 8582   1: 7032   Min.   : 0.00   1:23424
##  1st Qu.:31.00   M:54695   2:11794   2: 5204   1st Qu.: 5.00   2:12721
##  Median :44.00             3:12722   3:18905   Median :12.00   3:28403
##  Mean   :42.42             4:24816   4:12378   Mean   :12.54
##  3rd Qu.:52.00             5: 2377   5:11816   3rd Qu.:16.00
##  Max.   :92.00             6: 3884   6: 8407   Max.   :99.00
##                            7:  373   7:  806
##      duration          antskad          skadkost      motorage_gr
##  Min.   : 0.0000   Min.   :0.0000   Min.   :     0   1: 5104
##  1st Qu.: 0.4630   1st Qu.:0.0000   1st Qu.:     0   2: 8989
##  Median : 0.8274   Median :0.0000   Median :     0   3:50455
##  Mean   : 1.0107   Mean   :0.0108   Mean   :   264
##  3rd Qu.: 1.0000   3rd Qu.:0.0000   3rd Qu.:     0
##  Max.   :31.3397   Max.   :2.0000   Max.   :365347
##
```

In this data set we have 6 rating factors given by agarald (age of driver),which goes from 0 to 99, kon (sex of driver), either M (male) or K (female), zon (geographical zone, going from 1 to 7, generally going from more to less urban), mcklass (class of the vehicle, with 7 possible classes, as determined by the EV ratio of engine power to vehicle weight), fordald (age of vehicle, a numerical value between 0 and 99), bonuskl (bonus class based on experience, going from 1 to 7: a new driver starts with bonus class 1; for each claim-free year the bonus class is increased by 1; after the first claim the bonus is decreased by 2; the driver can not return to class 7 with less than 6 consecutive claim free years). There are 64548 observations in this data set, with one for each driver observed. We see that for some observations the duration is 0. The exposure (or duration) is determined using policy-years, i.e., how long the contract was in effect for each driver. For each observation we also have the number of claims (antskad) and the claim cost (skadkost).

Question 2

Use only the rating factors contained in the current tariff. Compute the exposure (in policy years), the claim frequency and the average claim severity for each rating factor (i.e., for each possible value of each rating factor).

Solution:

```
motorcycle <- data.frame(rating.factor = c(rep("Zone", nlevels(dataOhlsson$zon)),
                                           rep("Vehicle class", nlevels(dataOhlsson$mcklass)),
                                           rep("Vehicle age", nlevels(dataOhlsson$motorage_gr)),
                                           rep("Bonus class", nlevels(dataOhlsson$bonuskl))),
                         class = c(levels(dataOhlsson$zon),
                                   levels(dataOhlsson$mcklass),
                                   levels(dataOhlsson$motorage_gr),
                                   levels(dataOhlsson$bonuskl)),
                         stringsAsFactors = FALSE)

new.cols <- foreach (rating.factor = c("zon","mcklass","motorage_gr","bonuskl"), .combine = rbind) %do% {
                nclaims <- tapply(dataOhlsson$antskad,dataOhlsson[[rating.factor]], sum)
                sums <- tapply(dataOhlsson$duration, dataOhlsson[[rating.factor]],sum)
                severity <- tapply(dataOhlsson$skadkost,dataOhlsson[[rating.factor]],sum)
                n.levels <- nlevels(dataOhlsson[[rating.factor]])
                contrasts(dataOhlsson[[rating.factor]]) <-
                    contr.treatment(n.levels)[rank(-sums, ties.method = "first"), ]
                data.frame(duration = sums, n.claims = nclaims, totcostperm = severity/1000)}

motorcycle <- cbind(motorcycle, new.cols)
rm(new.cols)
```

After grouping the classes for the vehicle age into 3 groups (0-2, 3-5, 6 and up) and the bonus class into three groups (1-2,3-4,5-7) we get

```
motorcycle
```

```
##    rating.factor class   duration n.claims totcostperm
## 1           Zone     1  6205.3096      183    5539.963
## 2           Zone     2 10103.0904      167    4811.166
## 3           Zone     3 11676.5726      123    2522.628
## 4           Zone     4 32628.4931      196    3774.629
## 5           Zone     5  1582.1123        9     104.739
## 6           Zone     6  2799.9452       18     288.045
## 7           Zone     7   241.2877        1       0.650
## 11 Vehicle class     1  5190.3507       46     993.062
## 21 Vehicle class     2  3990.1151       57     883.137
## 31 Vehicle class     3 21665.6794      166    5371.543
## 41 Vehicle class     4 11739.8821       98    2191.578
## 51 Vehicle class     5 13439.9260      149    3297.119
## 61 Vehicle class     6  8880.1342      175    4160.776
## 71 Vehicle class     7   330.7233        6     144.605
## 12   Vehicle age     1  4955.4027      126    4964.419
## 22   Vehicle age     2  9753.8109      145    5506.945
## 32   Vehicle age     3 50527.5972      426    6570.456
## 13   Bonus class     1 19893.3698      207    4558.072
## 23   Bonus class     2  9615.7644      121    3627.142
## 33   Bonus class     3 35727.6766      369    8856.606
```

Question 3

Use a relative Poisson glm to determine relativities for the claim frequency, using the current rating factors. Provide a 95% confidence interval for each relativity. Comment on the overall fit of this model to the data.

Solution:

```
# Group the data by rating factors
grDataOhlsson <- dataOhlsson %>% group_by(bonuskl, zon, mcklass, motorage_gr) %>% summarize(duration = sum(durati
on),
                                                                                             antskad=sum(antskad),
                                                                                             skadkost=sum(skadkos
t))
```

```
## `summarise()` has grouped output by 'bonuskl', 'zon', 'mcklass'. You can
## override using the `.groups` argument.
```

```
# Fit the Poisson GLM
model.freq <- glm(antskad ~ zon + mcklass + motorage_gr + bonuskl + offset(log(duration)),
                  family = poisson,
                  data = grDataOhlsson[grDataOhlsson$duration>0,])

# Examine model fit
with(model.freq, cbind(res.deviance = deviance,
                       df = df.residual,
                       p = pchisq(deviance, df.residual, lower.tail = FALSE)))
```

```
##      res.deviance  df        p
## [1,]     360.2168 389 0.849545
```

With a deviance of 360.2168 on 389 degrees of freedom, the model seems a reasonable fit (p-value of 0.8495).

```
# Generate the table for relative frequency
P.zon = contrasts(dataOhlsson$zon)
P.mcklass = contrasts(dataOhlsson$mcklass)
P.motorage_gr = contrasts(dataOhlsson$motorage_gr)
P.bonuskl = contrasts(dataOhlsson$bonuskl)

mat.freq = summary(model.freq)$coefficients[, 1:2]

table.q3 = matrix(0, nrow = dim(motorcycle)[1] + 1, ncol = 5)
colnames(table.q3) = c("Rating factor", "Class", "Multiplier", "Lower bound", "Upper bound")

# Intercept
estimate = mat.freq[1, 1]
std.error = mat.freq[1, 2]
table.q3[1, ] = c("Intercept", "0", exp(estimate), exp(estimate - 1.96 * std.error), exp(estimate + 1.96 * std.er
ror))

# Zone
estimate = P.zon %*% mat.freq[2:7, 1]
std.error = P.zon %*% mat.freq[2:7, 2]
table.q3[2:8, ] = cbind(rep("Zone", 7), seq(7), exp(estimate), exp(estimate - 1.96 * std.error), exp(estimate +
1.96 * std.error))

# Vehicle class
estimate = P.mcklass %*% mat.freq[8:13, 1]
std.error = P.mcklass %*% mat.freq[8:13, 2]
table.q3[9:15, ] = cbind(rep("Vehicle class", 7), seq(7), exp(estimate), exp(estimate - 1.96 * std.error), exp(es
timate + 1.96 * std.error))

# Vehicle age
estimate = P.motorage_gr %*% mat.freq[14:15, 1]
std.error = P.motorage_gr %*% mat.freq[14:15, 2]
table.q3[16:18, ] = cbind(rep("Vehicle age", 3), seq(3), exp(estimate), exp(estimate - 1.96 * std.error), exp(est
imate + 1.96 * std.error))

# Bonus class
estimate = P.motorage_gr %*% mat.freq[16:17, 1]
std.error = P.motorage_gr %*% mat.freq[16:17, 2]
table.q3[19:21, ] = cbind(rep("Bonus class", 3), seq(3), exp(estimate), exp(estimate - 1.96 * std.error), exp(est
imate + 1.96 * std.error))

as_tibble(table.q3)
```

```
## # A tibble: 21 × 5
##    `Rating factor` Class Multiplier          `Lower bound`       `Upper bound`
##    <chr>           <chr> <chr>               <chr>               <chr>
##  1 Intercept       0     0.00234497033954873 0.00185823280920728 0.002959201810…
##  2 Zone            1     5.15619166509121    4.20563293521715    6.321596034815
##  3 Zone            2     2.72512290204893    2.21581010710497    3.351503275239…
##  4 Zone            3     1.70851751432824    1.36352962622445    2.1407911061302
##  5 Zone            4     1                   1                   1
##  6 Zone            5     0.906778337752537   0.464785438953093   1.769089314995…
##  7 Zone            6     1.03510019180549    0.638608104454922   1.677761994565…
##  8 Zone            7     0.727879994146364   0.102010643107309   5.193666756136…
##  9 Vehicle class   1     1.47808347047319    1.06238984535683    2.056430372743…
## 10 Vehicle class   2     2.1033504745518     1.5548677882858     2.845311512739…
## 11 Vehicle class   3     1                   1                   1
## 12 Vehicle class   4     1.32127812054859    1.02781661273447    1.698528560650…
## 13 Vehicle class   5     2.04515054171011    1.63104519770581    2.564392908388…
## 14 Vehicle class   6     3.97983541065762    3.18692242660292    4.970026808216…
## 15 Vehicle class   7     3.31183418327132    1.46441692364757    7.489838092122…
## 16 Vehicle age     1     3.23993950825312    2.64393379410273    3.970299120406…
## 17 Vehicle age     2     1.89477011838352    1.56371868235732    2.295907724340…
## 18 Vehicle age     3     1                   1                   1
## 19 Bonus class     1     1.44301072819081    1.17184971216712    1.776917244637…
## 20 Bonus class     2     1.27596653236492    1.06785602203578    1.524634930289…
## 21 Bonus class     3     1                   1                   1
```

Note that I also accepted answers based on the quasi-Poisson family where a dispersion parameter is estimated, causing the CIs to be different from the above.

Question 4

Use a Gamma glm (with log link function) to determine relativities for the severity, still using the current rating factors. Provide a 95% confidence interval for each relativity. Comment on the overall fit of this model to the data.

Solution:

```
model.sev <- glm(skadkost/antskad ~ zon + mcklass + motorage_gr + bonuskl,
                 family=Gamma("log"),
                 data=grDataOhlsson[grDataOhlsson$skadkost>0,],
                 weights=antskad)

with(model.sev, cbind(res.deviance = deviance,
                      df = df.residual,
                      p = pchisq(deviance, df.residual, lower.tail = FALSE)))
```

```
##      res.deviance  df           p
## [1,]     351.1129 164 1.13842e-15
```

The severity model doesn't seem to be a very good fit. We reject the hypothesis that the data comes from this model based on the residual deviance, given by 351 on 164 degrees of freedom, as its corresponding p-value is 1.13842×10^{-15}.

```
# Generate the table for relative severity
mat.sev = summary(model.sev)$coefficients[, 1:2]

table.q4 = matrix(0, nrow = dim(motorcycle)[1] + 1, ncol = 5)
colnames(table.q4) = c("Rating factor", "Class", "Multiplier", "Lower bound", "Upper bound")

# Intercept
estimate = mat.sev[1, 1]
std.error = mat.sev[1, 2]
table.q4[1, ] = c("Intercept", "0", exp(estimate), exp(estimate - 1.96 * std.error), exp(estimate + 1.96 * std.er
ror))

# Zone
estimate = P.zon %*% mat.sev[2:7, 1]
std.error = P.zon %*% mat.sev[2:7, 2]
table.q4[2:8, ] = cbind(rep("Zone", 7), seq(7), exp(estimate), exp(estimate - 1.96 * std.error), exp(estimate +
1.96 * std.error))

# Vehicle class
estimate = P.mcklass %*% mat.sev[8:13, 1]
std.error = P.mcklass %*% mat.sev[8:13, 2]
table.q4[9:15, ] = cbind(rep("Vehicle class", 7), seq(7), exp(estimate), exp(estimate - 1.96 * std.error), exp(es
timate + 1.96 * std.error))

# Vehicle age
estimate = P.motorage_gr %*% mat.sev[14:15, 1]
std.error = P.motorage_gr %*% mat.sev[14:15, 2]
table.q4[16:18, ] = cbind(rep("Vehicle age", 3), seq(3), exp(estimate), exp(estimate - 1.96 * std.error), exp(est
imate + 1.96 * std.error))

# Bonus class
estimate = P.motorage_gr %*% mat.sev[16:17, 1]
std.error = P.motorage_gr %*% mat.sev[16:17, 2]
table.q4[19:21, ] = cbind(rep("Bonus class", 3), seq(3), exp(estimate), exp(estimate - 1.96 * std.error), exp(est
imate + 1.96 * std.error))

as_tibble(table.q4)
```

```
## # A tibble: 21 × 5
##    `Rating factor` Class Multiplier         `Lower bound`       `Upper bound`
##    <chr>           <chr> <chr>              <chr>               <chr>
##  1 Intercept       0     15698.1679330244   11322.3990123379    21765.0407996487
##  2 Zone            1     1.30039469226393   0.967902679007292   1.74710370406512
##  3 Zone            2     1.36973027703226   1.0186593887144     1.84179427648203
##  4 Zone            3     0.936380318134884  0.677075531329196   1.2949930393572
##  5 Zone            4     1                  1                   1
##  6 Zone            5     0.96339135135138   0.36576430540324    2.53748898443062
##  7 Zone            6     0.78451446232571   0.387815300888558   1.58700015236133
##  8 Zone            7     0.0176535454001371 0.00104756310507183 0.2974977485230…
##  9 Vehicle class   1     0.745942949078776  0.466065835461491   1.19388901941155
## 10 Vehicle class   2     0.667289538253253  0.430200054122025   1.0350424729048
## 11 Vehicle class   3     1                  1                   1
## 12 Vehicle class   4     0.797635610340983  0.555995256341655   1.14429495508695
## 13 Vehicle class   5     0.833034093463106  0.601004157341868   1.15464392782422
## 14 Vehicle class   6     1.03466202092935   0.750353624581498   1.42669464434278
## 15 Vehicle class   7     1.4329900900771    0.435415448857646   4.71609494712838
## 16 Vehicle age     1     2.55576215878475   1.91010115168082    3.41967241186585
## 17 Vehicle age     2     2.3454838512553    1.77392394528234    3.10120087793493
## 18 Vehicle age     3     1                  1                   1
## 19 Bonus class     1     1.03083084936552   0.766438737547944   1.38642814871705
## 20 Bonus class     2     0.835563065106195  0.645526021612162   1.08154530165342
## 21 Bonus class     3     1                  1                   1
```

Question 5

Assess whether rating factors for the policyholder's age and sex would have a significant impact. Include an interaction term. You should group the age factor into intervals of 0-30 and 30-100.

Solution:

```
dataOhlsson$driverage_gr <- cut(dataOhlsson$agarald,
                                breaks = c(0, 30,100),
                                labels = c("1", "2"),
                                right = FALSE)

motorcycle.new <-
    data.frame(rating.factor = c(rep("Bonus class", nlevels(dataOhlsson$bonuskl)),
                                 rep("Vehicle class", nlevels(dataOhlsson$mcklass)),
                                 rep("Vehicle age", nlevels(dataOhlsson$motorage_gr)),
                                 rep("Zone", nlevels(dataOhlsson$zon)),
                                 rep("Driver age", nlevels(dataOhlsson$driverage_gr)),
                                 rep("Sex", nlevels(dataOhlsson$kon))),
               class = c(levels(dataOhlsson$bonuskl),
                         levels(dataOhlsson$mcklass),
                         levels(dataOhlsson$motorage_gr),
                         levels(dataOhlsson$zon),
                         levels(dataOhlsson$driverage_gr),
                         levels(dataOhlsson$kon)),
               stringsAsFactors = FALSE)

new.cols <- foreach (rating.factor = c("zon", "mcklass", "motorage_gr", "bonuskl", "driverage_gr", "kon"), .combi
ne = rbind) %do% {
    nclaims <- tapply(dataOhlsson$antskad,dataOhlsson[[rating.factor]], sum)
    sums <- tapply(dataOhlsson$duration, dataOhlsson[[rating.factor]],sum)
    severity <- tapply(dataOhlsson$skadkost,dataOhlsson[[rating.factor]],sum)
    n.levels <- nlevels(dataOhlsson[[rating.factor]])
    contrasts(dataOhlsson[[rating.factor]]) <-
        contr.treatment(n.levels)[rank(-sums, ties.method = "first"), ]
    data.frame(duration = sums, n.claims = nclaims, sev = severity/nclaims)}

motorcycle.new <- cbind(motorcycle.new, new.cols)
rm(new.cols)

grDataOhlsson.new <- dataOhlsson %>% group_by(bonuskl, zon, mcklass, motorage_gr, kon, driverage_gr) %>% summariz
e(duration = sum(duration),

antskad=sum(antskad),

skadkost=sum(skadkost))
```

```
## `summarise()` has grouped output by 'bonuskl', 'zon', 'mcklass', 'motorage_gr',
## 'kon'. You can override using the `.groups` argument.
```

```
# Fit the Poisson GLM for frequency
model.freq.new <- glm(antskad ~ zon + mcklass + motorage_gr + bonuskl + kon * driverage_gr +offset(log(duratio
n)),
                     family = poisson,
                     data = grDataOhlsson.new[grDataOhlsson.new$duration>0,])

with(model.freq.new,
     cbind(res.deviance = deviance, df = df.residual,
     p = pchisq(deviance, df.residual, lower.tail = FALSE)))
```

```
##      res.deviance   df p
## [1,]     742.7633 1277 1
```

```
pchisq(model.freq.new$deviance - model.freq$deviance,
       model.freq.new$df.residual - model.freq$df.residual,
       lower.tail=FALSE)
```

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

We have combined the ages into 2 groups given by the breaks 0, 30, 100. When including age and sex with an interaction term, for the frequency the deviance goes to 742.76 on 1277 degrees of freedom. The LRT statistic we get $742.76 - 360.22 = 382.55$ on $1277 - 389 = 888$ degrees of freedom. The corresponding p-value is 1, suggesting that the simpler model is a better fit.

```
# Fit the Gamma GLM for severity
model.sev.new <- glm(skadkost/antskad ~ zon + mcklass + motorage_gr + bonuskl + kon * driverage_gr,
                     family = Gamma("log"),
                     data = grDataOhlsson.new[grDataOhlsson.new$antskad>0,],
                     weights = antskad)

with(model.sev.new,
     cbind(res.deviance = deviance, df = df.residual,
     p = pchisq(deviance, df.residual, lower.tail = FALSE)))
```

```
##      res.deviance  df            p
## [1,]     612.1475 288 1.706859e-25
```

```
pchisq(model.sev.new$deviance - model.sev$deviance * summary(model.sev)$dispersion / summary(model.sev.new)$dispe
rsion,
       model.sev.new$df.residual - model.sev$df.residual,
       lower.tail=FALSE)
```

```
## [1] 7.085095e-11
```

For the severity however, the LRT statistic is given by 253.12 on 124 degrees of freedom, with corresponding p-value given by 7.09×10^{-11}. Hence in this case, the age and sex do appear to provide a model that has a better fit. We also see that for both the frequency and the severity models, while sex is not a significant factor, the interaction term between age and sex and the coefficient for the age group 0-30 are both significant.

Question 6

Now combine your multiplier estimates for the frequency and severity data to get multipliers (and associated 95% confidence intervals) for the premium overall and then propose a new tariff based on your analysis. Compare the results to the old tariff.

Solution:

```r
# Generate the table for new tariff
table.q6 = matrix(0, nrow = dim(motorcycle)[1] + 1, ncol = 5)
colnames(table.q6) = c("Rating factor", "Class", "Multiplier", "Lower bound", "Upper bound")

# Intercept
estimate.freq = mat.freq[1, 1]
std.error.freq = mat.freq[1, 2]
estimate.sev = mat.sev[1, 1]
std.error.sev = mat.sev[1, 2]
table.q6[1, ] = c("Intercept", "0", exp(estimate.freq + estimate.sev),
                  exp(estimate.freq + estimate.sev - 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)),
                  exp(estimate.freq + estimate.sev + 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)))

# Zone
estimate.freq = P.zon %*% mat.freq[2:7, 1]
std.error.freq = P.zon %*% mat.freq[2:7, 2]
estimate.sev = P.zon %*% mat.sev[2:7, 1]
std.error.sev = P.zon %*% mat.sev[2:7, 2]
table.q6[2:8, ] = cbind(rep("Zone", 7), seq(7), exp(estimate.freq + estimate.sev),
                  exp(estimate.freq + estimate.sev - 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)),
                  exp(estimate.freq + estimate.sev + 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)))

# Vehicle class
estimate.freq = P.mcklass %*% mat.freq[8:13, 1]
std.error.freq = P.mcklass %*% mat.freq[8:13, 2]
estimate.sev = P.mcklass %*% mat.sev[8:13, 1]
std.error.sev = P.mcklass %*% mat.sev[8:13, 2]
table.q6[9:15, ] = cbind(rep("Vehicle class", 7), seq(7), exp(estimate.freq + estimate.sev),
                   exp(estimate.freq + estimate.sev - 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)),
                   exp(estimate.freq + estimate.sev + 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)))
# Vehicle age
estimate.freq = P.motorage_gr %*% mat.freq[14:15, 1]
std.error.freq = P.motorage_gr %*% mat.freq[14:15, 2]
estimate.sev = P.motorage_gr %*% mat.sev[14:15, 1]
std.error.sev = P.motorage_gr %*% mat.sev[14:15, 2]
table.q6[16:18, ] = cbind(rep("Vehicle age", 3), seq(3), exp(estimate.freq + estimate.sev),
                    exp(estimate.freq + estimate.sev - 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)),
                    exp(estimate.freq + estimate.sev + 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)))
# Bonus class
estimate.freq = P.motorage_gr %*% mat.freq[16:17, 1]
std.error.freq = P.motorage_gr %*% mat.freq[16:17, 2]
estimate.sev = P.motorage_gr %*% mat.sev[16:17, 1]
std.error.sev = P.motorage_gr %*% mat.sev[16:17, 2]
table.q6[19:21, ] = cbind(rep("Bonus class", 3), seq(3), exp(estimate.freq + estimate.sev),
                    exp(estimate.freq + estimate.sev - 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)),
                    exp(estimate.freq + estimate.sev + 1.96 * sqrt(std.error.freq^2 + std.error.sev^2)))

# Append the old tariff
tariff.old <- c(0,
                7.768, 4.227, 1.336, 1.000, 1.734, 1.402, 1.402,
                0.625, 0.769, 1.000, 1.406, 1.875, 4.062, 6.873,
                2.000, 1.200, 1.000,
                1.250, 1.125, 1.000)

table.q6 <- cbind(table.q6, tariff.old)
as_tibble(table.q6)
```

```
## # A tibble: 21 × 6
##    `Rating factor` Class Multiplier       `Lower bound` `Upper bound` tariff.old
##    <chr>           <chr> <chr>            <chr>         <chr>         <chr>
##  1 Intercept       0     36.8117381881973 24.648022842… 54.978205638… 0
##  2 Zone            1     6.70508427358011 4.6837045832… 9.5988451698… 7.768
##  3 Zone            2     3.73268334757043 2.6009695305… 5.3568197588… 4.227
##  4 Zone            3     1.5998221736057  1.0777977596… 2.3746857555… 1.336
##  5 Zone            4     1                1             1             1
##  6 Zone            5     0.8735824081835… 0.2693243878… 2.8335578147… 1.734
##  7 Zone            6     0.81205107042752 0.3456349663… 1.9078710351… 1.402
##  8 Zone            7     0.0128496625225… 0.0004116919… 0.4010616354… 1.402
##  9 Vehicle class   1     1.10256594294937 0.6206206667… 1.9587676073… 0.625
## 10 Vehicle class   2     1.40354376694843 0.8237368936… 2.3914615465… 0.769
## 11 Vehicle class   3     1                1             1             1
## 12 Vehicle class   4     1.05389848011396 0.6789605646… 1.6358858882… 1.406
## 13 Vehicle class   5     1.70368012750906 1.1452046573… 2.5345041676… 1.875
## 14 Vehicle class   6     4.11778454895722 2.7862232716… 6.0857109924… 4.062
## 15 Vehicle class   7     4.74582556460639 1.1200092081… 20.109531356… 6.873
## 16 Vehicle age     1     8.280514791945   5.8053034789… 11.811083687… 2
## 17 Vehicle age     2     4.44415271450964 3.1665581941… 6.2372115523… 1.2
## 18 Vehicle age     3     1                1             1             1
## 19 Bonus class     1     1.48749997458448 1.0355565557… 2.1366830832… 1.25
## 20 Bonus class     2     1.06615050675576 0.7792268847… 1.4587239292… 1.125
## 21 Bonus class     3     1                1             1             1
```

We see that the new tariff is much higher than the old one for Vehicle Age 1 and 2. We also see that for Zone 5 to 7, the new tariff suggests to lower the premium compared to the baseline of Zone 4, while for the old tariff they were all higher than the baseline. The number of claims for these 3 zones is very small though, so it seems like it might be best to not make such an important change based on such a small sample.

Question 7

Comment on any further analysis that should be considered before deciding on a final tariff.

Solution:

we saw that, especially for the severity, the age and sex seem to be significant factors that should be considered. On the other hand, looking at the results for coefficients in each of the frequency and severity models (reproduced below), we see that Zones 5,6,7 (note that although the results below are based on the ordering of levels by decreasing order of duration, and therefore do not necessarily correspond to the original numbering, for these 3 particular classes are actually the same as in the original ordering) do not seem to be significant, and should therefore probably be combined with the baseline of Zone 4. Finally, we should explore the use of models other than the gamma distribution for the severity, to see if a better fit can be obtained. 5