47956712VoKhuongDuyAssignment1

2024-04-24

```
set.seed(10)
```

```
install.packages("moments",repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/khuon/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'moments' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\khuon\AppData\Local\Temp\RtmpiG6QvJ\downloaded_packages
```

```
library(boot)
library(moments)
```

Question 1

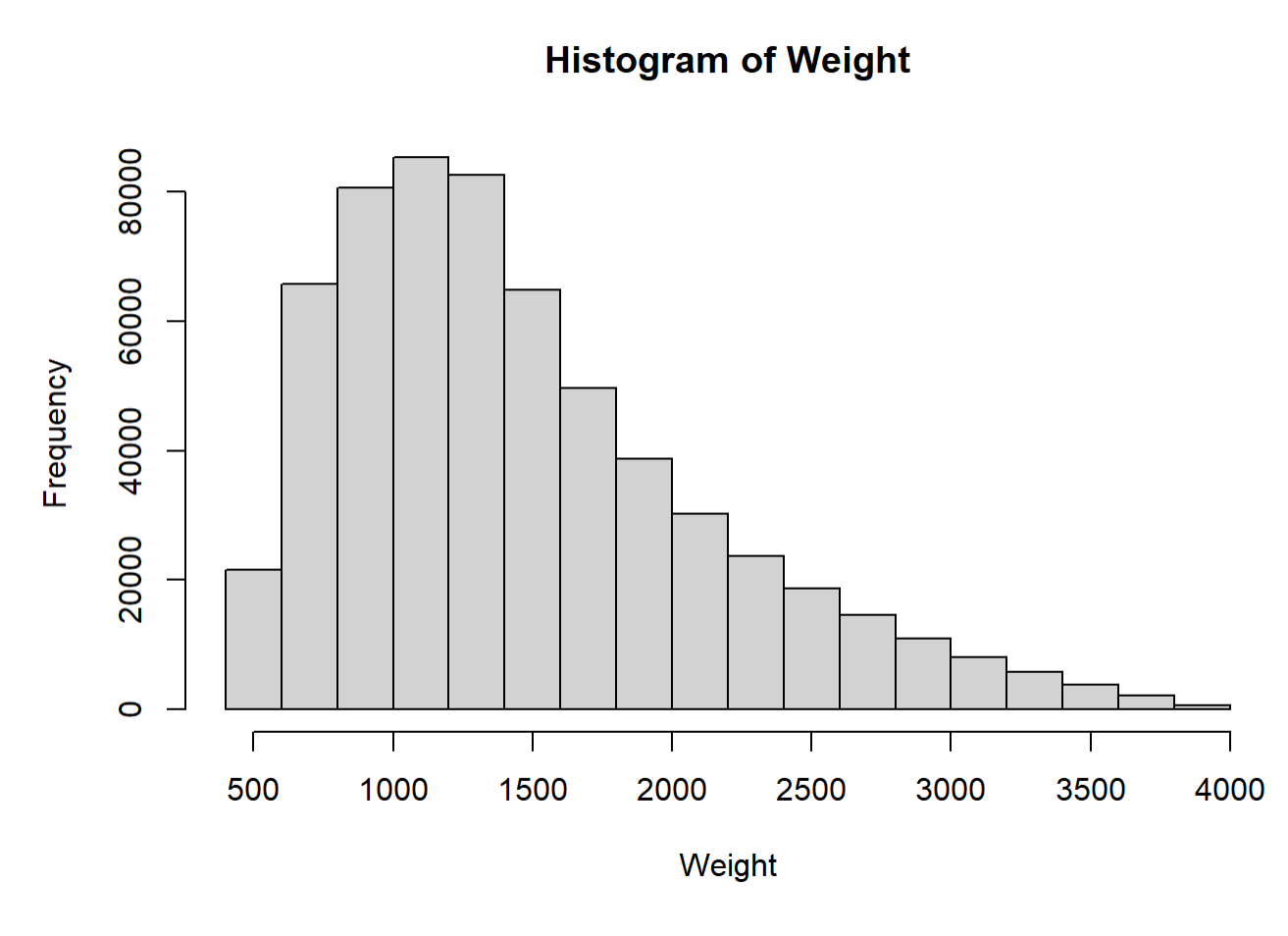
```
Assignment1_Dataset_2024 <- read.csv("Assignment1_Dataset_2024.csv")
attach(Assignment1_Dataset_2024)
summary(Assignment1_Dataset_2024)
```

```
##      Counts           exposure         distance         weight
##  Min.   :0.00000   Min.   :0.8000   Min.   : 1.00   Min.   : 450
##  1st Qu.:0.00000   1st Qu.:0.8500   1st Qu.: 5.00   1st Qu.: 962
##  Median :0.00000   Median :0.8999   Median :10.00   Median :1319
##  Mean   :0.06092   Mean   :0.8999   Mean   :14.85   Mean   :1464
##  3rd Qu.:0.00000   3rd Qu.:0.9498   3rd Qu.:19.00   3rd Qu.:1826
##  Max.   :3.00000   Max.   :1.0000   Max.   :95.00   Max.   :3994
##       age            carage          state              gender
##  Min.   :18.00   Min.   : 1.000   Length:607697      Length:607697
##  1st Qu.:35.00   1st Qu.: 3.000   Class :character   Class :character
##  Median :46.00   Median : 5.000   Mode  :character   Mode  :character
##  Mean   :47.25   Mean   : 7.762
##  3rd Qu.:58.00   3rd Qu.:10.000
##  Max.   :98.00   Max.   :45.000
```

```
Assignment1_Dataset_2024$state <- as.factor(Assignment1_Dataset_2024$state)
Assignment1_Dataset_2024$gender <- as.factor(Assignment1_Dataset_2024$gender)
summary(Assignment1_Dataset_2024)
```

```
##      Counts           exposure         distance         weight
##  Min.   :0.00000   Min.   :0.8000   Min.   : 1.00   Min.   : 450
##  1st Qu.:0.00000   1st Qu.:0.8500   1st Qu.: 5.00   1st Qu.: 962
##  Median :0.00000   Median :0.8999   Median :10.00   Median :1319
##  Mean   :0.06092   Mean   :0.8999   Mean   :14.85   Mean   :1464
##  3rd Qu.:0.00000   3rd Qu.:0.9498   3rd Qu.:19.00   3rd Qu.:1826
##  Max.   :3.00000   Max.   :1.0000   Max.   :95.00   Max.   :3994
##       age            carage        state          gender
##  Min.   :18.00   Min.   : 1.000   ACT:120952   female:243052
##  1st Qu.:35.00   1st Qu.: 3.000   NSW:121441   male  :364645
##  Median :46.00   Median : 5.000   QLD:121990
##  Mean   :47.25   Mean   : 7.762   SA :122092
##  3rd Qu.:58.00   3rd Qu.:10.000   VIC:121222
##  Max.   :98.00   Max.   :45.000
```

```
hist(weight, xlab="Weight", main="Histogram of Weight")
```




```
skewness(weight)
```

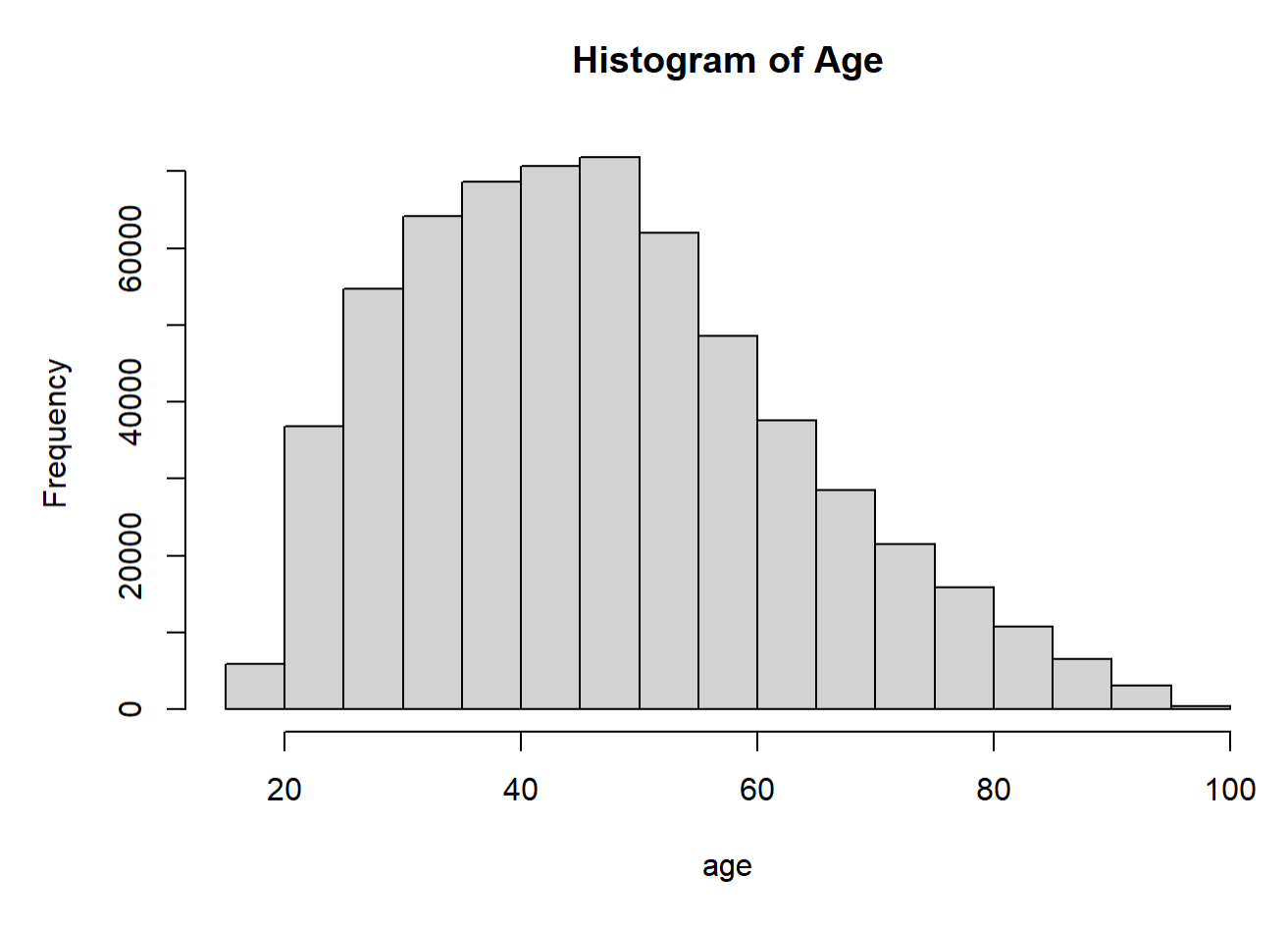
```
## [1] 0.9708882
```

```
hist(distance, xlab="Distance", main="Histogram of Distance")
```


```
skewness(distance)
```

```
## [1] 1.943979
```

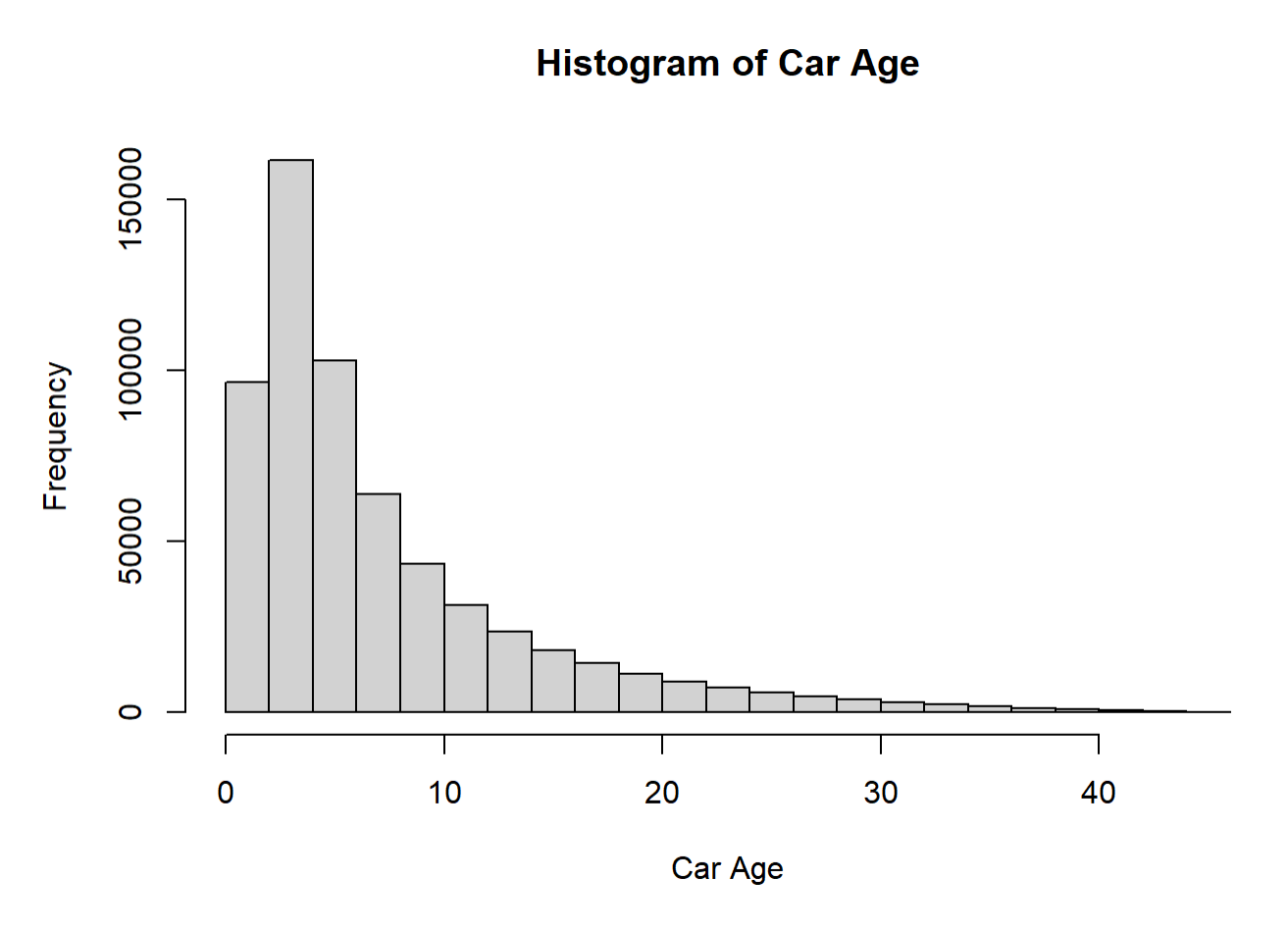
```
hist(age, xlab="age", main="Histogram of Age")
```




```
skewness(age)
```

```
## [1] 0.5068129
```

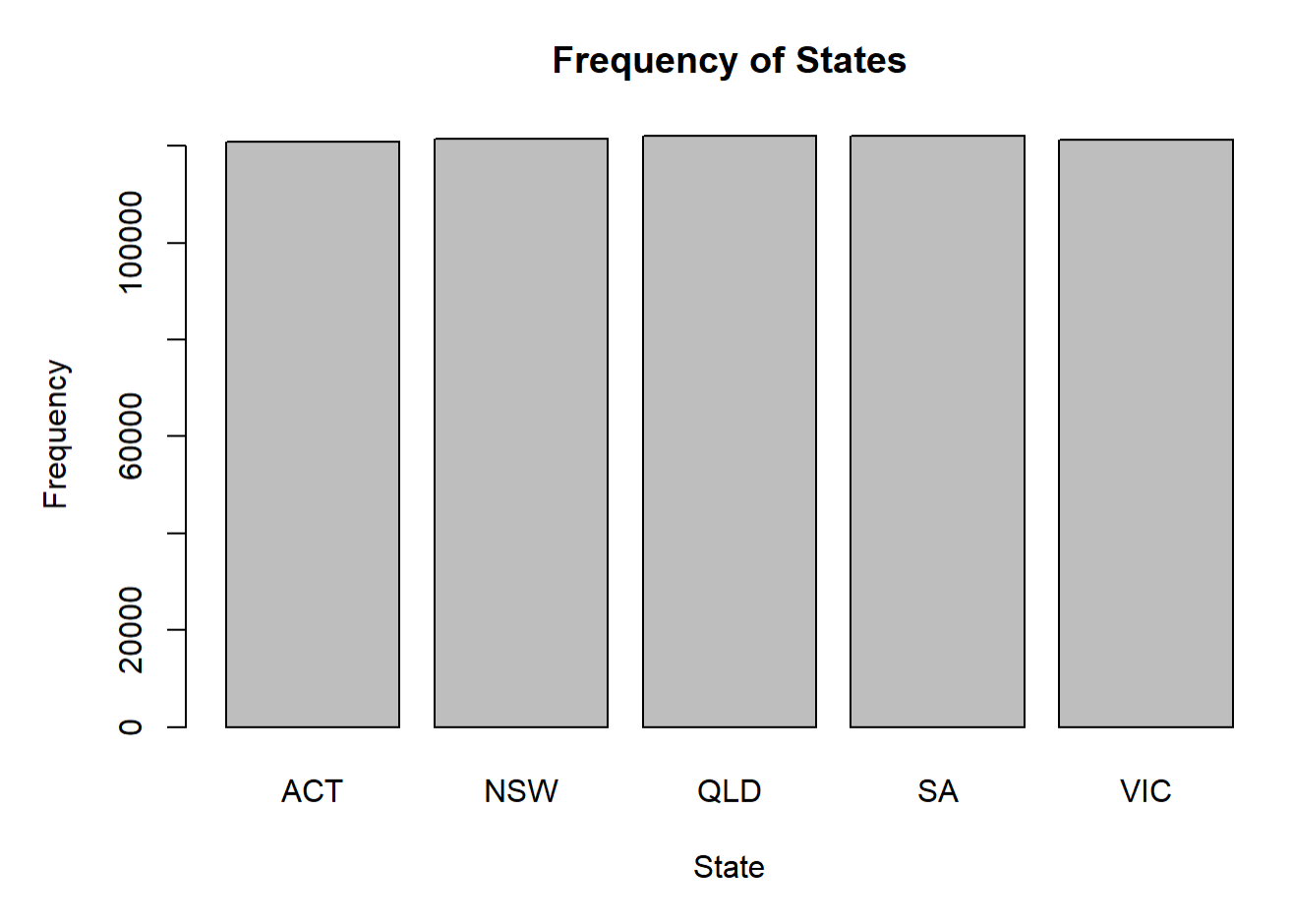
```
hist(carage, xlab="Car Age", main="Histogram of Car Age")
```



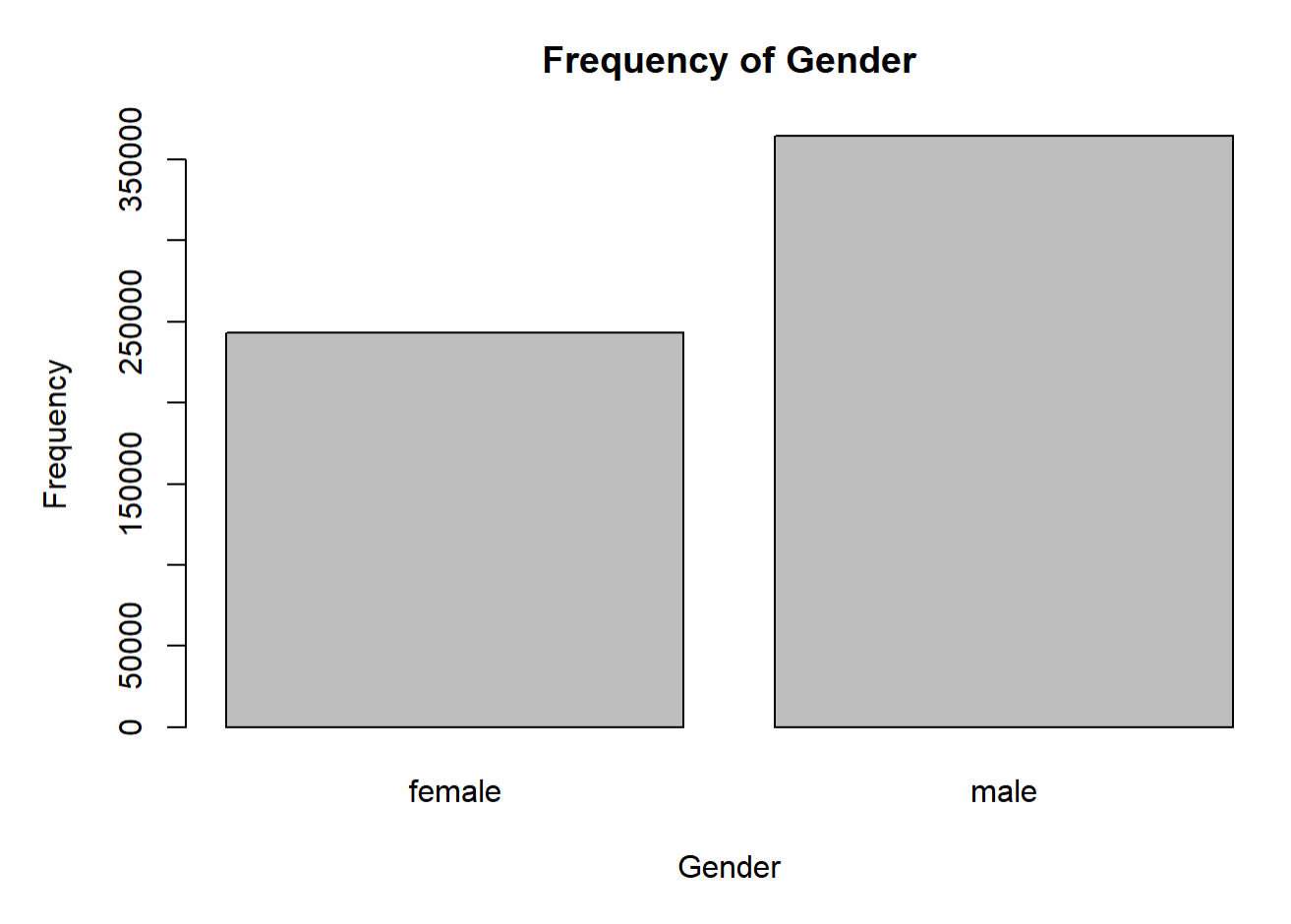

```
skewness(carage)
```

```
## [1] 1.91042
```

```
barplot(table(state),xlab="State", ylab="Frequency", main="Frequency of States")
```




```
barplot(table(gender), xlab="Gender", ylab="Frequency", main="Frequency of Gender")
```



Question 2

```
model1_full <- glm(Counts~weight+distance+age+carage+gender+state, data=Assignment1_Dataset_2024, family=poisson
(link=log), offset=log(exposure))
summary(model1_full)
```

```
##
## Call:
## glm(formula = Counts ~ weight + distance + age + carage + gender +
##     state, family = poisson(link = log), data = Assignment1_Dataset_2024,
##     offset = log(exposure))
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -2.681e+00  2.430e-02 -110.335   <2e-16 ***
## weight       1.894e-04  7.339e-06   25.806   <2e-16 ***
## distance     3.923e-03  3.401e-04   11.534   <2e-16 ***
## age         -8.988e-03  3.366e-04  -26.699   <2e-16 ***
## carage       2.074e-02  6.579e-04   31.533   <2e-16 ***
## gendermale  -1.686e-01  1.047e-02  -16.104   <2e-16 ***
## stateNSW    -6.404e-03  1.634e-02   -0.392   0.6951
## stateQLD    -2.634e-02  1.640e-02   -1.606   0.1084
## stateSA     -1.620e-02  1.636e-02   -0.990   0.3221
## stateVIC    -3.908e-02  1.648e-02   -2.372   0.0177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 210184  on 607696  degrees of freedom
## Residual deviance: 207505  on 607687  degrees of freedom
## AIC: 279989
##
## Number of Fisher Scoring iterations: 6
```

```
model1_red <- glm(Counts~weight+distance+age+carage+gender, data=Assignment1_Dataset_2024, family=poisson(link=lo
g), offset=log(exposure))
summary(model1_red)
```

```
##
## Call:
## glm(formula = Counts ~ weight + distance + age + carage + gender,
##     family = poisson(link = log), data = Assignment1_Dataset_2024,
##     offset = log(exposure))
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.698e+00  2.201e-02 -122.60   <2e-16 ***
## weight       1.894e-04  7.339e-06   25.80   <2e-16 ***
## distance     3.922e-03  3.401e-04   11.53   <2e-16 ***
## age         -8.988e-03  3.366e-04  -26.70   <2e-16 ***
## carage       2.074e-02  6.579e-04   31.53   <2e-16 ***
## gendermale  -1.686e-01  1.047e-02  -16.11   <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: 210184  on 607696  degrees of freedom
## Residual deviance: 207512  on 607691  degrees of freedom
## AIC: 279988
##
## Number of Fisher Scoring iterations: 6
```

```
LR <- model1_red$deviance - model1_full$deviance
LR
```

```
## [1] 7.184716
```

```
qchisq(p = .05, df = 4, lower.tail = FALSE)
```

```
## [1] 9.487729
```

```
model1 <- model1_red
coefficients(model1)
```

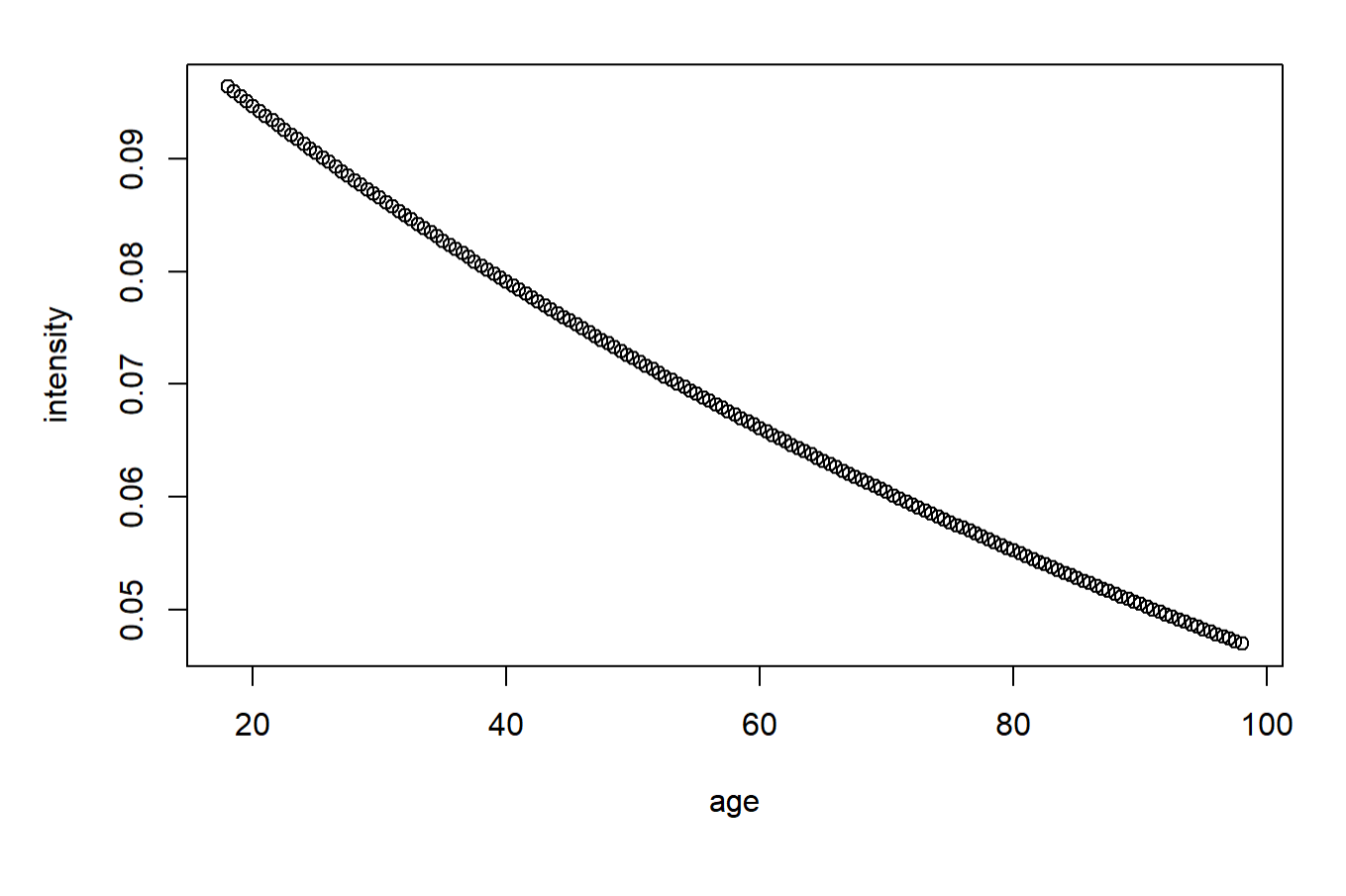
```
##   (Intercept)        weight      distance           age        carage
## -2.6982576416  0.0001893555  0.0039215852 -0.0089879197  0.0207449284
##    gendermale
## -0.1686028685
```

```
new_data = data.frame(weight=2000, distance=15, age=30, carage=4, gender="female", state="NSW", exposure=1)
predict(model1,new_data,type="response")
```

```
##          1
## 0.08651991
```

```
xage_1 <- seq(min(age),max(age),0.5)

yage_1 <- predict(model1, list(age=xage_1,
                weight=rep(2000,length(xage_1)),
                distance=rep(15, length(xage_1)),
                age=rep(30,length(xage_1)),
                carage=rep(4,length(xage_1)),
                gender=rep("female",length(xage_1)),
                state=rep("NSW",length(xage_1)),
                exposure=rep(1, length(xage_1))),
type="response")
plot(xage_1,yage_1,xlab="age",ylab="intensity")
```



Question 3:

```
model2_full <- glm(Counts~weight+distance+age+carage+gender+state+I(weight^2)+I(distance^2)+I(age^2)+I(carage^2),
data=Assignment1_Dataset_2024, family=poisson(link=log), offset=log(exposure))
summary(model2_full)
```

```
##
## Call:
## glm(formula = Counts ~ weight + distance + age + carage + gender +
##     state + I(weight^2) + I(distance^2) + I(age^2) + I(carage^2),
##     family = poisson(link = log), data = Assignment1_Dataset_2024,
##     offset = log(exposure))
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.303e+00  5.197e-02 -44.309  < 2e-16 ***
## weight         1.826e-04  3.274e-05   5.576 2.45e-08 ***
## distance       1.966e-03  9.633e-04   2.040   0.0413 *
## age           -2.581e-02  1.714e-03 -15.063  < 2e-16 ***
## carage         2.458e-02  2.064e-03  11.905  < 2e-16 ***
## gendermale    -1.685e-01  1.047e-02 -16.092  < 2e-16 ***
## stateNSW      -6.786e-03  1.634e-02  -0.415   0.6779
## stateQLD      -2.635e-02  1.640e-02  -1.607   0.1081
## stateSA       -1.624e-02  1.636e-02  -0.993   0.3208
## stateVIC      -3.922e-02  1.648e-02  -2.380   0.0173 *
## I(weight^2)    1.799e-09  8.585e-09   0.210   0.8340
## I(distance^2)  3.262e-05  1.498e-05   2.177   0.0295 *
## I(age^2)       1.687e-04  1.681e-05  10.040  < 2e-16 ***
## I(carage^2)   -1.242e-04  6.359e-05  -1.953   0.0508 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 210184  on 607696  degrees of freedom
## Residual deviance: 207399  on 607683  degrees of freedom
## AIC: 279890
##
## Number of Fisher Scoring iterations: 6
```

```
step(model2_full, direction="backward")
```

```
## Start:  AIC=279890.5
## Counts ~ weight + distance + age + carage + gender + state +
##     I(weight^2) + I(distance^2) + I(age^2) + I(carage^2)
##
##                 Df Deviance    AIC
## - I(weight^2)    1   207399 279888
## - state          4   207406 279890
## <none>               207399 279890
## - I(carage^2)    1   207402 279892
## - distance       1   207403 279893
## - I(distance^2)  1   207403 279893
## - weight         1   207430 279920
## - I(age^2)       1   207496 279986
## - carage         1   207542 280032
## - age            1   207618 280108
## - gender         1   207656 280145
##
## Step:  AIC=279888.5
## Counts ~ weight + distance + age + carage + gender + state +
##     I(distance^2) + I(age^2) + I(carage^2)
##
##                 Df Deviance    AIC
## - state          4   207406 279888
## <none>               207399 279888
## - I(carage^2)    1   207402 279890
## - distance       1   207403 279891
## - I(distance^2)  1   207403 279891
## - I(age^2)       1   207496 279984
## - carage         1   207542 280030
## - age            1   207618 280106
## - gender         1   207656 280143
## - weight         1   208038 280526
##
## Step:  AIC=279887.7
## Counts ~ weight + distance + age + carage + gender + I(distance^2) +
##     I(age^2) + I(carage^2)
##
##                 Df Deviance    AIC
## <none>               207406 279888
## - I(carage^2)    1   207410 279889
## - distance       1   207410 279890
## - I(distance^2)  1   207410 279890
## - I(age^2)       1   207504 279984
## - carage         1   207549 280029
## - age            1   207625 280105
## - gender         1   207663 280143
## - weight         1   208045 280525
```

```
##
## Call:  glm(formula = Counts ~ weight + distance + age + carage + gender +
##     I(distance^2) + I(age^2) + I(carage^2), family = poisson(link = log),
##     data = Assignment1_Dataset_2024, offset = log(exposure))
##
## Coefficients:
##   (Intercept)         weight       distance            age         carage
##    -2.326e+00      1.892e-04      1.963e-03     -2.581e-02      2.458e-02
##    gendermale  I(distance^2)       I(age^2)    I(carage^2)
##    -1.685e-01      3.265e-05      1.687e-04     -1.241e-04
##
## Degrees of Freedom: 607696 Total (i.e. Null);  607688 Residual
## Null Deviance:       210200
## Residual Deviance: 207400    AIC: 279900
```

```
model2 <- glm(formula=Counts~weight+distance+age+carage+gender+I(distance^2)+I(age^2)+I(carage^2),data=Assignment
1_Dataset_2024, family=poisson(link=log), offset=log(exposure))
summary(model2)
```

```
##
## Call:
## glm(formula = Counts ~ weight + distance + age + carage + gender +
##     I(distance^2) + I(age^2) + I(carage^2), family = poisson(link = log),
##     data = Assignment1_Dataset_2024, offset = log(exposure))
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.326e+00  4.463e-02 -52.106   <2e-16 ***
## weight         1.892e-04  7.339e-06  25.781   <2e-16 ***
## distance       1.963e-03  9.633e-04   2.037   0.0416 *
## age           -2.581e-02  1.714e-03 -15.063   <2e-16 ***
## carage         2.458e-02  2.064e-03  11.905   <2e-16 ***
## gendermale    -1.685e-01  1.047e-02 -16.094   <2e-16 ***
## I(distance^2)  3.265e-05  1.498e-05   2.180   0.0293 *
## I(age^2)       1.687e-04  1.681e-05  10.040   <2e-16 ***
## I(carage^2)   -1.241e-04  6.358e-05  -1.952   0.0509 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 210184  on 607696  degrees of freedom
## Residual deviance: 207406  on 607688  degrees of freedom
## AIC: 279888
##
## Number of Fisher Scoring iterations: 6
```

```
coefficients(model2)
```

```
##   (Intercept)        weight      distance           age        carage
## -2.3256445193  0.0001891998  0.0019627489 -0.0258140249  0.0245759909
##    gendermale I(distance^2)      I(age^2)   I(carage^2)
## -0.1684686178  0.0000326505  0.0001687494 -0.0001241414
```

```
predict(model2,new_data,type="response")
```

```
##          1
## 0.08745019
```

```
xage_2 <- seq(min(age),max(age),0.5)
yage_2 <- predict(model2, list(age=xage_2,
                weight=rep(2000,length(xage_2)),
                distance=rep(15, length(xage_2)),
                age=rep(30,length(xage_2)),
                carage=rep(4,length(xage_2)),
                gender=rep("female",length(xage_2)),
                state=rep("NSW",length(xage_2)),
                exposure=rep(1, length(xage_2))),
type="response")
plot(xage_2,yage_2,xlab="age",ylab="intensity", col="red", ylim=c(0.04,0.12))
points(xage_1,yage_1,col="blue")
```

Question 4

```
LR_2 <- model1$deviance - model2$deviance
LR_2
```

```
## [1] 106.6293
```

```
qchisq(p = .05, df = 3, lower.tail = FALSE)
```

```
## [1] 7.814728
```

Question 5

```
cv.error_model1=cv.glm(Assignment1_Dataset_2024,model1,K=10)$delta[1]
cv.error_model2=cv.glm(Assignment1_Dataset_2024,model2,K=10)$delta[1]
cv.error_model1
```

```
## [1] 0.06072492
```

```
cv.error_model2
```

```
## [1] 0.06071381
```