ACST3061 Assignment

Yunbae Chae

2023-04-27

Contents

Question 1

```
library(dplyr)
```

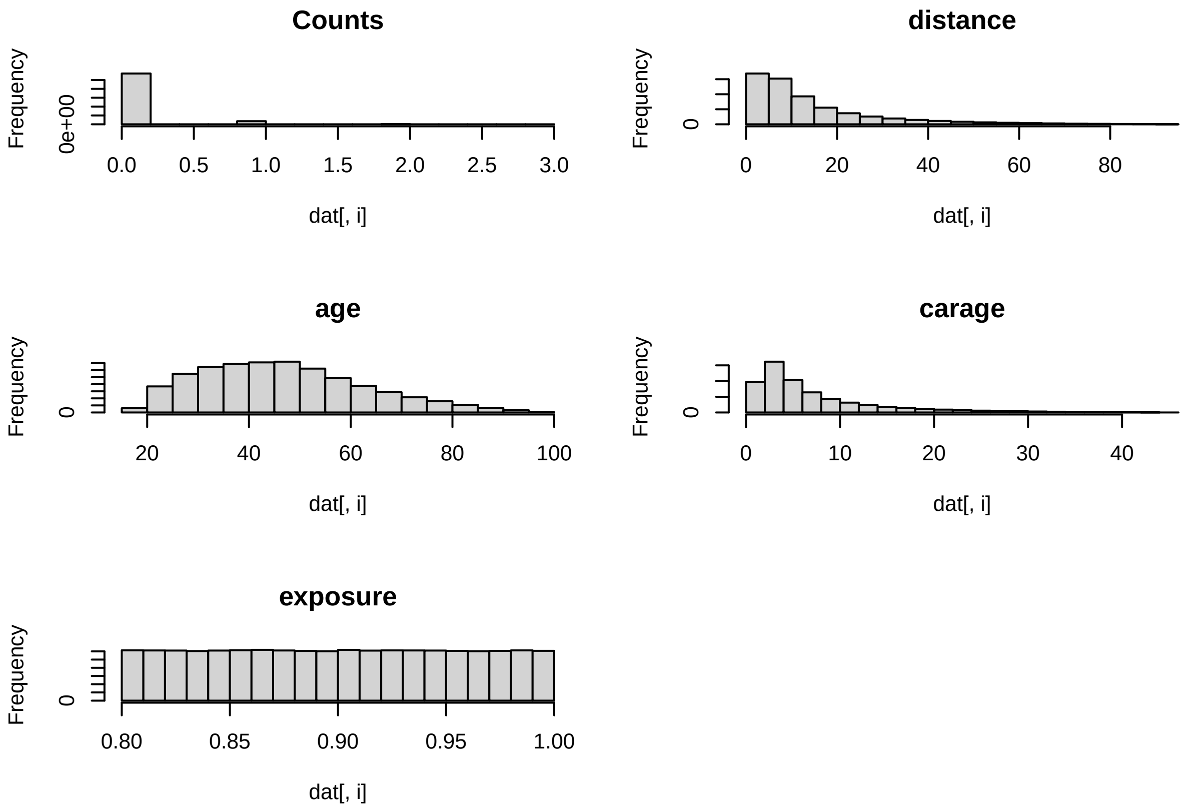
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
dat <- read.csv("Dataset2023.csv")
dat <- data.frame(dat)
par(mfrow=c(3,2))
for (i in c(1,3,4,5,6)) {
  hist(dat[,i], main = colnames(dat)[i])
}
summary(dat)
```

```
##      Counts            gender             distance          age
##  Min.   :0.00000   Length:609499      Min.   : 1.00   Min.   :18.00
##  1st Qu.:0.00000   Class :character   1st Qu.: 5.00   1st Qu.:35.00
##  Median :0.00000   Mode  :character   Median :10.00   Median :46.00
##  Mean   :0.06094                      Mean   :14.85   Mean   :47.25
##  3rd Qu.:0.00000                      3rd Qu.:19.00   3rd Qu.:58.00
##  Max.   :3.00000                      Max.   :95.00   Max.   :98.00
##      carage          exposure
##  Min.   : 1.000   Min.   :0.8000
##  1st Qu.: 3.000   1st Qu.:0.8500
##  Median : 5.000   Median :0.8999
##  Mean   : 7.762   Mean   :0.8999
##  3rd Qu.:10.000   3rd Qu.:0.9498
##  Max.   :45.000   Max.   :1.0000
```



The data seem to be right-skewed so it can be expected there will be some extreme values. The Counts variable has few outliers and roughly 96% of insured have not claimed an insurance. The book consists of approximately 60% male and 40% female clients.

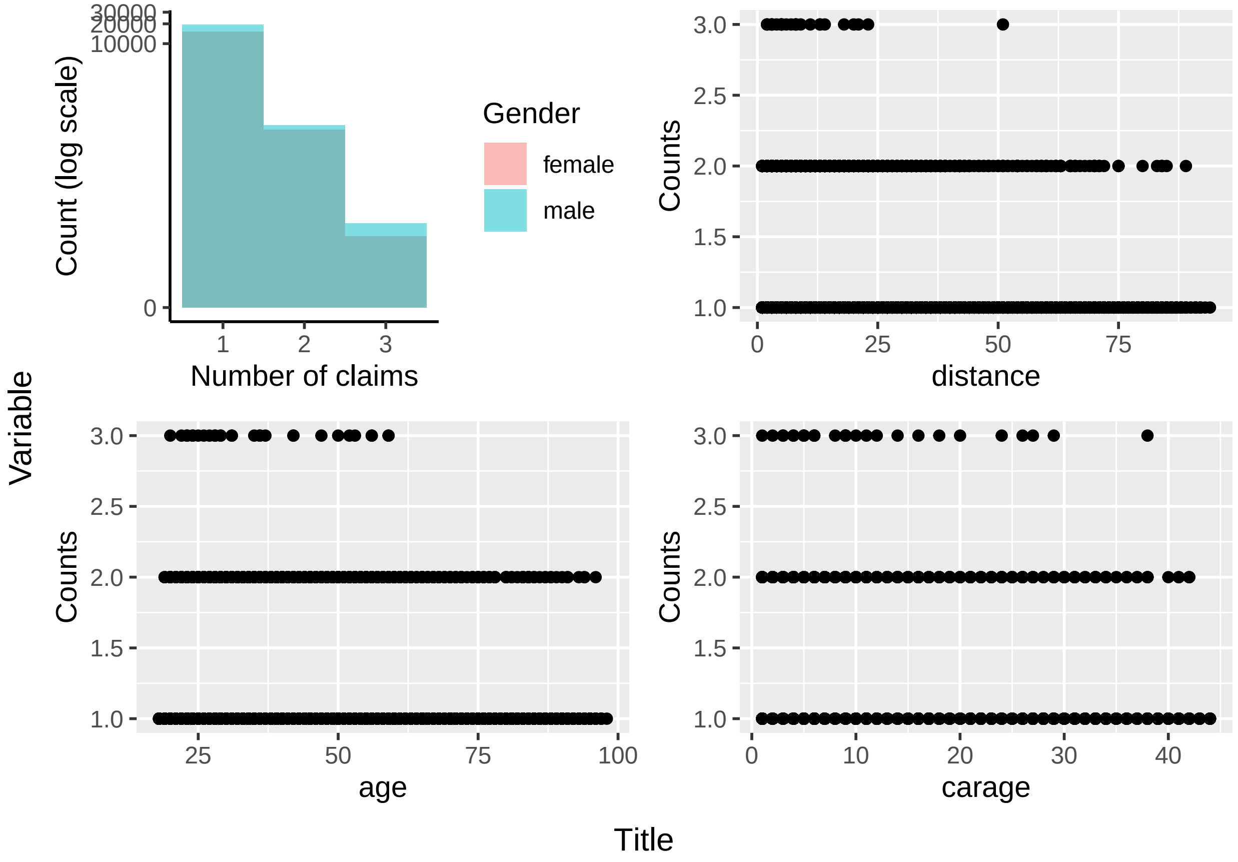
```
datno0 <- filter(dat,Counts>0)

Countsdata <- data.frame(Counts=count(filter(dat,Counts>0)),"1"=count(filter(dat,Counts==1)),"2"=count(
colnames(Countsdata)[1] <- "Counts"
colnames(Countsdata)[2] <- "1"
colnames(Countsdata)[3] <- "2"
colnames(Countsdata)[4] <- "3"
Countsdata
```

```
##   Counts     1    2  3
## 1  36005 34899 1077 29
```

```
gender <- ggplot(datno0,aes(x=Counts,fill=gender))+geom_histogram(position="identity",binwidth=1,alpha=
distance <- ggplot(datno0,aes(x=distance,y=Counts))+geom_point()
age <- ggplot(datno0,aes(x=age,y=Counts))+geom_point()
carage <- ggplot(datno0,aes(x=carage,y=Counts))+geom_point()
par(mfrow=c(1,1))
plots <- list(gender, distance, age, carage)

gridExtra::grid.arrange(
  grobs = lapply(plots, ggplotGrob),
  nrow = 2,
  ncol = 2,
  bottom = "Title",
  left = "Variable"
)
```



‘0 Counts’ entries were removed from graphing due to computing power insufficiency (Rendering error due to too many number of entries). For distance, age and car age variables, it is difficult to judge any visual implications, except only to notice that higher claim counts are more concentrated in the lower distance, age and car age. This must be due to the histograms displayed above being right-skewed. In the end, there seems to be no clear visual patterns that stand out.

Question 2

```
model1 <- glm(Counts~gender+distance+age+carage,family=poisson(link=log),data=dat,offset=log(exposure))
summary(model1)
```

```
##
## Call:
## glm(formula = Counts ~ gender + distance + age + carage, family = poisson(link = log),
##     data = dat, offset = log(exposure))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6175  -0.3669  -0.3387  -0.3121   4.3732
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.4112425  0.0187803 -128.39   <2e-16 ***
## gendermale  -0.1677929  0.0104510  -16.05   <2e-16 ***
## distance     0.0039102  0.0003395   11.52   <2e-16 ***
## age         -0.0090343  0.0003362  -26.87   <2e-16 ***
## carage       0.0208245  0.0006564   31.72   <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: 210844  on 609498  degrees of freedom
## Residual deviance: 208797  on 609494  degrees of freedom
## AIC: 281506
##
## Number of Fisher Scoring iterations: 6
```

The estimated regression coefficients for gender, distance, age and carage are -0.1677929, 0.0039102, -0.0090343 and 0.0208245 respectively.

The result shows all covariates under the 5% p-value level so they are all statistically significant. Each covariates are tested if they are statistically non-different to 0 using the Wald test:

$$H_0 : \beta_{ij} = 0,\ H_1 : \beta_{ij} \neq 0$$

$$W = \frac{\hat{\beta_{ij}} - \beta_{ij}}{s.e.(\beta_{ij})}$$

Equivalent results displayed in the above report as ‘z value’.

$$p\ value = P(Z > |W|)$$

These are also displayed above under ‘Pr(>|z|)’.
The ‘p value’ less than the 5% threshold indicates enough evidence to reject the null hypothesis. In the above case, all four covariates show very small p-values therefore all of their H_0’s are rejected; They are all different to 0. They are all significant.

Question 3

```
model2 <- glm(Counts~gender+distance+carage,family=poisson(link=log),data=dat,offset=log(exposure))
summary(model2)
```

```
##
## Call:
## glm(formula = Counts ~ gender + distance + carage, family = poisson(link = log),
##     data = dat, offset = log(exposure))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5988  -0.3611  -0.3387  -0.3201   4.3645
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.8280082  0.0111175 -254.37   <2e-16 ***
## gendermale  -0.1679959  0.0104510  -16.07   <2e-16 ***
## distance     0.0039285  0.0003395   11.57   <2e-16 ***
## carage       0.0208196  0.0006566   31.71   <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: 210844  on 609498  degrees of freedom
## Residual deviance: 209539  on 609495  degrees of freedom
## AIC: 282246
##
## Number of Fisher Scoring iterations: 6
```

The estimated regression coefficients for Model 2 for gender, distance and carage are -0.1679959, 0.0039285 and 0.0208196 respectively.

To conduct the likelihood ratio test:

$$H_0 : lnL_{full} - lnL_{reduced} = 0,\ \ H_1 : notH_0$$

we need the scaled deviance for the reduced and the full model. Since the dispersion parameter is 1 for the Poisson distribution, scaled deviance is the same as the deviance (residual deviance).

From the R output, the residual deviance for the reduced model: 209700 and for the full model: 208959.

Then the likelihood ratio test statistic is 209539-208797 (=742).

The full model estimates 4 parameters whereas the reduced model estimates 3. Then the quantile of the χ^2_p must be based on 1 degree of freedom.

$$\chi^2_1 = 3.841459$$

Since 742>3.841459, there is a strong evidence to reject the null hypothesis and conclude that one model is statistically better than the other.

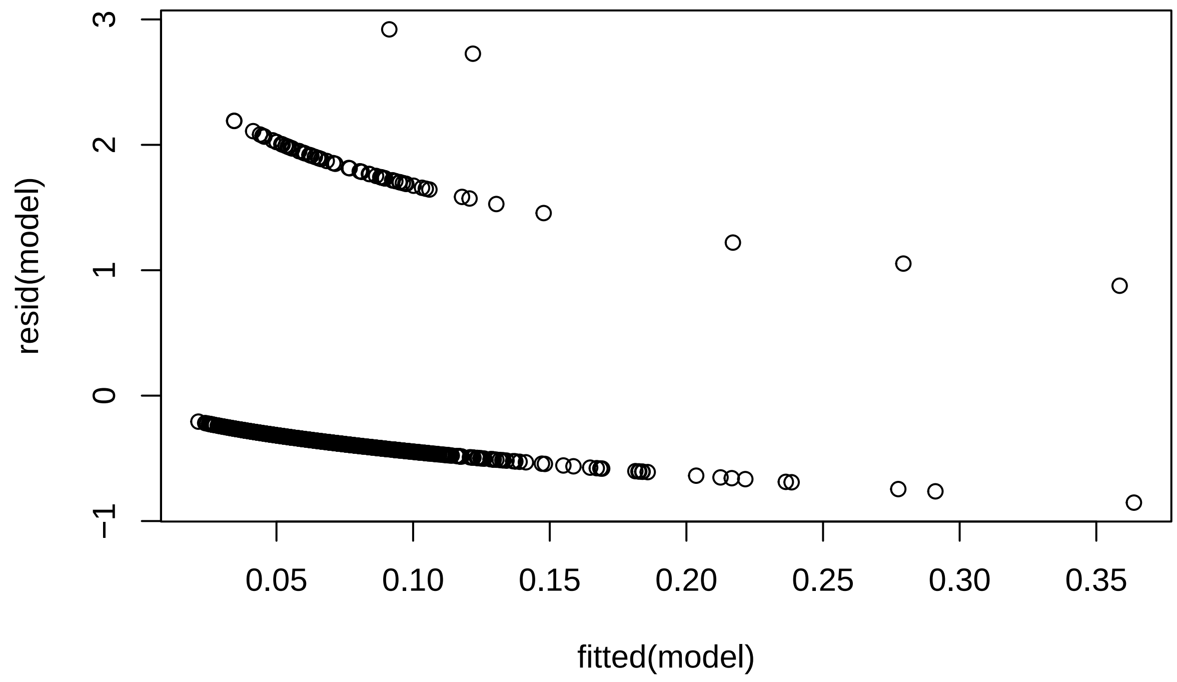
Comparing their residual deviances; 208797 for Model 1 and 209539 for Model 2, Model 1 has the lower residual deviance so Model 1 is statistically better than Model 2.

Question 4

```
dat2 <- head(dat,1000)
model <- glm(Counts~gender+distance+age+carage,family=poisson(link=log),data=dat2,offset=log(exposure))
model
```

```
##
## Call:  glm(formula = Counts ~ gender + distance + age + carage, family = poisson(link = log),
##     data = dat2, offset = log(exposure))
##
## Coefficients:
## (Intercept)   gendermale     distance          age       carage
##   -1.983194    -0.245070     0.003095    -0.017514     0.036128
##
## Degrees of Freedom: 999 Total (i.e. Null);  995 Residual
## Null Deviance:       382
## Residual Deviance: 368.8     AIC: 518
```

```
plot(resid(model)~fitted(model))
```



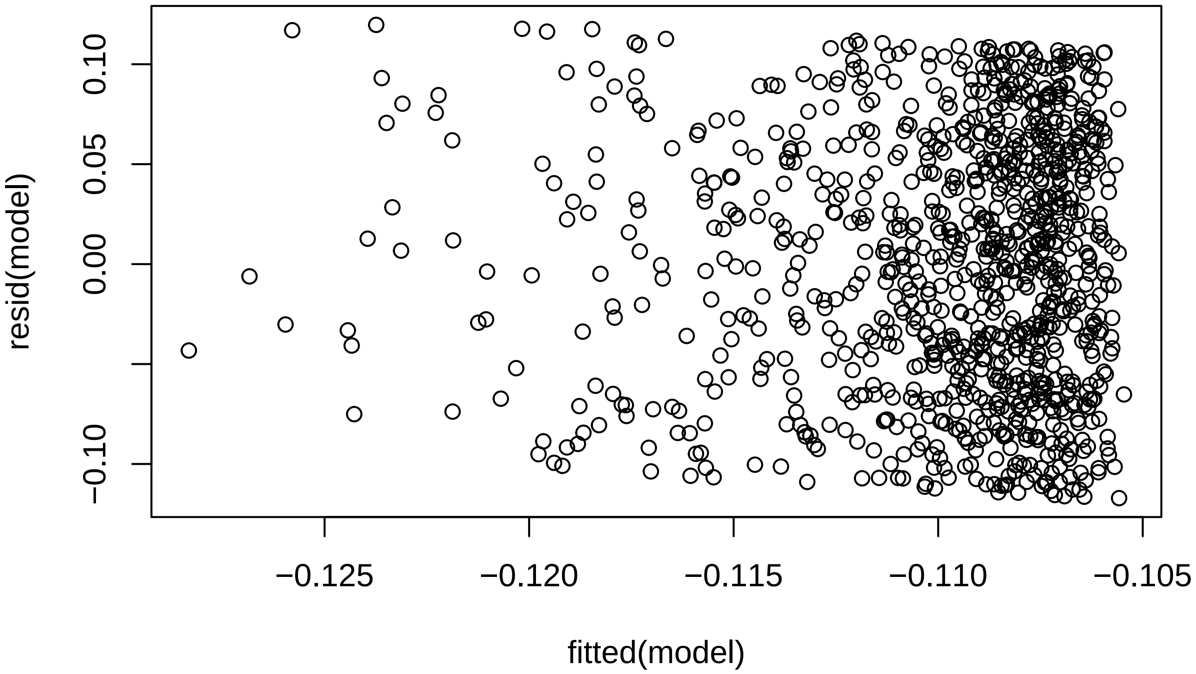
The residuals vs fitted plot shows 3 horizontally (or mild hint of exponentially) distributed points that are not scattered around 0. The clear presence of pattern and non-randomness of distribution of points lead to the conclusion that the model is not a good fit.

Question 5

```r
dat2 <- head(dat,1000)
model <- glm(log(exposure)~gender+distance+age+carage,data=dat2)
model
```

```
##
## Call:  glm(formula = log(exposure) ~ gender + distance + age + carage,
##     data = dat2)
##
## Coefficients:
## (Intercept)   gendermale     distance          age       carage
##  -1.061e-01    1.266e-03   -2.670e-04   -1.008e-05   -1.912e-05
##
## Degrees of Freedom: 999 Total (i.e. Null);  995 Residual
## Null Deviance:       3.99
## Residual Deviance: 3.976     AIC: -2678
```

```r
plot(resid(model)~fitted(model))
```



The residual plot shows a randomly distributed points around 0 without any obvious patterns. The model is a good fit.