Computer Intensive Methods - Final projects (2022)

Group : Deo Byabazaire (2159254)
Mirriam Dianah Lucheveleli (2159277)
Farida Iddy (2159270)
Quynh Long Khuong (2159280)

Contents

1 Project 1

```
library("DAAG")
data(nassCDS)
names(nassCDS)
```

```
##  [1] "dvcat"       "weight"      "dead"        "airbag"      "seatbelt"   
##  [6] "frontal"     "sex"         "ageOFocc"    "yearacc"     "yearVeh"    
## [11] "abcat"       "occRole"     "deploy"      "injSeverity" "caseid"
```

```
dim(nassCDS)
```

```
## [1] 26217    15
```

```
# Check missing value
sapply(nassCDS, function(x){sum(is.na(x))})
```

```
##       dvcat      weight        dead      airbag    seatbelt     frontal 
##           0           0           0           0           0           0 
##         sex    ageOFocc     yearacc     yearVeh       abcat     occRole 
##           0           0           0           1           0           0 
##      deploy injSeverity      caseid 
##           0         153           0
```

```
# complete-case data
nassCDS <- na.omit(nassCDS)
dim(nassCDS)
```

```
## [1] 26063    15
```

1.1 Question 1

Let Y_i be an indicator variable which takes the value of 1 if an occupant died in an accident (the variable dead) and zero otherwise and X_i be the age of occupant in years (the variable ageOFocc). We consider the following GLM

$$g(P(Y_i = 1)) = \beta_0 + \beta_1 X_i$$

1. Estimate the model using the classical GLM approach

```
nassCDS %<>% mutate(dead = ifelse(dead == "dead", 1, 0))
glm_dead <- glm(dead ~ ageOFocc, data = nassCDS, family = "binomial")
summary(glm_dead)
```

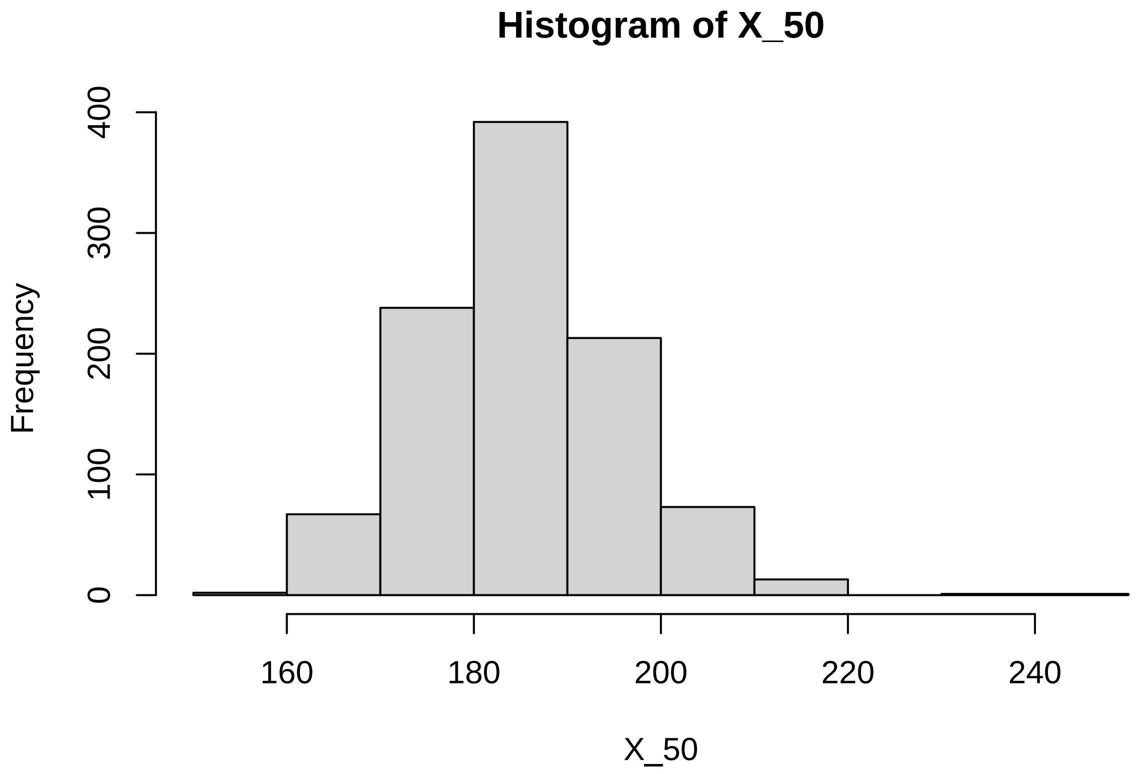
```
## 
## Call:
## glm(formula = dead ~ ageOFocc, family = "binomial", data = nassCDS)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.5396  -0.3220  -0.2757  -0.2484   2.6821
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.907983   0.072013  -54.27   <2e-16 ***
## ageOFocc     0.021183   0.001484   14.27   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9610.0  on 26062  degrees of freedom
## Residual deviance: 9418.2  on 26061  degrees of freedom
## AIC: 9422.2
##
## Number of Fisher Scoring iterations: 6
```

2. Let X_50 be the age of occupant for which the probability to die is 0.5 $P(Y_i = 1) = 0.5$. Estimate X_50. Use non parametric bootstrap to estimate the distribution of X_50 and construct a 95% for the X_50

```
B <- 1000
n <- length(nassCDS$dead)
index <- c(1:n)
X_50 <- c()

for (i in seq(B)) {
    index.b <- sample(index, n, replace=TRUE)
    dead <- nassCDS$dead[index.b]
    ageOFocc <- nassCDS$ageOFocc[index.b]
    glm_dead <- glm(dead ~ ageOFocc, family = "binomial")
    X_50[i] <- (-coef(glm_dead)[[1]])/coef(glm_dead)[[2]]
}
# Distribution of X50
hist(X_50)
```



```
# Estimate 95% CI
c(quantile(X_50, 0.025), quantile(X_50, 0.975))
```

```
##    2.5%   97.5%
## 166.541 207.007
```