stat431 a3 q4

Yiming Shen 20891774

19/11/2023

```r
# step 0
N <- 1000
x1 <- rnorm(N, 0, 1)
x2 <- rbinom(N, 1, 0.25)

estimates <- matrix(NA,nrow=3, ncol=500)
CIs <- matrix(NA,nrow=6,ncol=500)
```

(a)

```r
for (i in 1:500) {

  beta0 <- 0.5
  beta1 <- 0.2
  beta2 <- 0.5
  y <- rpois(N, exp(beta0 + beta1*x1 + beta2*x2))

  mod <- glm(y ~ x1 + x2, family = poisson)

  est_beta0 <- summary(mod)$coefficients[1,1]
  se0 <- summary(mod)$coefficients[1,2]

  est_beta1 <- summary(mod)$coefficients[2,1]
  se1 <- summary(mod)$coefficients[2,2]

  est_beta2 <- summary(mod)$coefficients[3,1]
  se2 <- summary(mod)$coefficients[3,2]

  estimates[1,i] <- est_beta0
  estimates[2,i] <- est_beta1
  estimates[3,i] <- est_beta2

  c <- 1.96
  CIs[1,i] <- est_beta0 - c*se0
  CIs[2,i] <- est_beta0 + c*se0
  CIs[3,i] <- est_beta1 - c*se1
  CIs[4,i] <- est_beta1 + c*se1
  CIs[5,i] <- est_beta2 - c*se2
  CIs[6,i] <- est_beta2 + c*se2
}

# average estimates
```

```r
average_est_beta0 <- mean(estimates[1,])
average_est_beta1 <- mean(estimates[2,])
average_est_beta2 <- mean(estimates[3,])

average_est_beta0
```

```
## [1] 0.4987404
```

```r
average_est_beta1
```

```
## [1] 0.1994879
```

```r
average_est_beta2
```

```
## [1] 0.5001897
```

```r
# proportion of confidence interval coverage
coverage <- matrix(NA, nrow=3, ncol=500)

for (i in 1:500) {
  L0 <- CIs[1,i]
  U0 <- CIs[2,i]
  L1 <- CIs[3,i]
  U1 <- CIs[4,i]
  L2 <- CIs[5,i]
  U2 <- CIs[6,i]

  coverage[1,i] <- (beta0 >= L0) & (U0 >= beta0)
  coverage[2,i] <- (beta1 >= L1) & (U1 >= beta1)
  coverage[3,i] <- (beta2 >= L2) & (U2 >= beta2)
}

beta0_pro <- mean(coverage[1,])
beta1_pro <- mean(coverage[2,])
beta2_pro <- mean(coverage[3,])

beta0_pro
```

```
## [1] 0.962
```

```r
beta1_pro
```

```
## [1] 0.956
```

```r
beta2_pro
```

```
## [1] 0.958
```

Comments: Based on the results, we found that average estimates are very close to the true values; the proportion of confidence intervals that cover true value is very close to 0.95 (since we constructed 95% CI).

(b)

```r
N <- 1000
x1 <- rnorm(N, 0, 1)
x2 <- rbinom(N, 1, 0.25)

estimates <- matrix(NA,nrow=3, ncol=500)
```

```r
CIs <- matrix(NA,nrow=6,ncol=500)


for (i in 1:500) {

  beta0 <- 0.5
  beta1 <- 0.2
  beta2 <- 0.5
  gamma0 <- 0.5
  gamma1 <- 0
  gamma2 <- 0.5

  y_star <- rpois(N, exp(beta0 + beta1*x1 + beta2*x2))

  pie <- exp(gamma0 + gamma1*x1 + gamma2*x2) /
    (1 + exp(gamma0 + gamma1*x1 + gamma2*x2))

  w <- rbinom(N, 1, pie)
  y <- c(NA, N)
  y[w==1] <- 0
  y[w==0] <- y_star[w==0]


  # fit model
  mod <- glm(y ~ x1 + x2, family = poisson)

  est_beta0 <- summary(mod)$coefficients[1,1]
  se0 <- summary(mod)$coefficients[1,2]

  est_beta1 <- summary(mod)$coefficients[2,1]
  se1 <- summary(mod)$coefficients[2,2]

  est_beta2 <- summary(mod)$coefficients[3,1]
  se2 <- summary(mod)$coefficients[3,2]

  estimates[1,i] <- est_beta0
  estimates[2,i] <- est_beta1
  estimates[3,i] <- est_beta2

  c <- 1.96
  CIs[1,i] <- est_beta0 - c*se0
  CIs[2,i] <- est_beta0 + c*se0
  CIs[3,i] <- est_beta1 - c*se1
  CIs[4,i] <- est_beta1 + c*se1
  CIs[5,i] <- est_beta2 - c*se2
  CIs[6,i] <- est_beta2 + c*se2

}

# average estimates
average_est_beta0 <- mean(estimates[1,])
average_est_beta1 <- mean(estimates[2,])
average_est_beta2 <- mean(estimates[3,])
```

```r
average_est_beta0
```

```
## [1] -0.4781659
```

```r
average_est_beta1
```

```
## [1] 0.1956083
```

```r
average_est_beta2
```

```
## [1] 0.1494656
```

```r
# proportion of confidence interval coverage
coverage <- matrix(NA, nrow=3, ncol=500)

for (i in 1:500) {
  L0 <- CIs[1,i]
  U0 <- CIs[2,i]
  L1 <- CIs[3,i]
  U1 <- CIs[4,i]
  L2 <- CIs[5,i]
  U2 <- CIs[6,i]

  coverage[1,i] <- (beta0 >= L0) & (U0 >= beta0)
  coverage[2,i] <- (beta1 >= L1) & (U1 >= beta1)
  coverage[3,i] <- (beta2 >= L2) & (U2 >= beta2)
}

beta0_pro <- mean(coverage[1,])
beta1_pro <- mean(coverage[2,])
beta2_pro <- mean(coverage[3,])

beta0_pro
```

```
## [1] 0
```

```r
beta1_pro
```

```
## [1] 0.792
```

```r
beta2_pro
```

```
## [1] 0.108
```

Comments: Based on the results, we found that average estimates are not close to the true values; the proportion of confidence intervals that cover true value is much lower than 0.95 (we constructed 95% CI).

(c)

```r
N <- 1000
x1 <- rnorm(N, 0, 1)
x2 <- rbinom(N, 1, 0.25)

estimates <- matrix(NA,nrow=3, ncol=500)
CIs <- matrix(NA,nrow=6,ncol=500)


for (i in 1:500) {
```

```
  beta0 <- 0.5
  beta1 <- 0.2
  beta2 <- 0.5
  gamma0 <- 0.5
  gamma1 <- 0
  gamma2 <- 0.5

  y_star <- rpois(N, exp(beta0 + beta1*x1 + beta2*x2))

  pie <- exp(gamma0 + gamma1*x1 + gamma2*x2) /
    (1 + exp(gamma0 + gamma1*x1 + gamma2*x2))

  w <- rbinom(N, 1, pie)
  y <- c(NA, N)
  y[w==1] <- 0
  y[w==0] <- y_star[w==0]


  # fit model
  require(pscl)
  mod <- zeroinfl(y ~ x1 + x2)

  est_beta0 <- summary(mod)$coefficients$count[1]
  se0 <- sqrt(summary(mod)$vcov[1,1])

  est_beta1 <- summary(mod)$coefficients$count[2]
  se1 <- sqrt(summary(mod)$vcov[2,2])

  est_beta2 <- summary(mod)$coefficients$count[3]
  se2 <- sqrt(summary(mod)$vcov[3,3])

  estimates[1,i] <- est_beta0
  estimates[2,i] <- est_beta1
  estimates[3,i] <- est_beta2

  c <- 1.96
  CIs[1,i] <- est_beta0 - c*se0
  CIs[2,i] <- est_beta0 + c*se0
  CIs[3,i] <- est_beta1 - c*se1
  CIs[4,i] <- est_beta1 + c*se1
  CIs[5,i] <- est_beta2 - c*se2
  CIs[6,i] <- est_beta2 + c*se2

}
```

```
## Loading required package: pscl
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

```r
# average estimates
average_est_beta0 <- mean(estimates[1,])
average_est_beta1 <- mean(estimates[2,])
average_est_beta2 <- mean(estimates[3,])

average_est_beta0
```

```
## [1] 0.4938926
```

```r
average_est_beta1
```

```
## [1] 0.1998101
```

```r
average_est_beta2
```

```
## [1] 0.502593
```

```r
# proportion of confidence interval coverage
coverage <- matrix(NA, nrow=3, ncol=500)

for (i in 1:500) {
  L0 <- CIs[1,i]
  U0 <- CIs[2,i]
  L1 <- CIs[3,i]
  U1 <- CIs[4,i]
  L2 <- CIs[5,i]
  U2 <- CIs[6,i]

  coverage[1,i] <- (beta0 >= L0) & (U0 >= beta0)
  coverage[2,i] <- (beta1 >= L1) & (U1 >= beta1)
  coverage[3,i] <- (beta2 >= L2) & (U2 >= beta2)
}

beta0_pro <- mean(coverage[1,])
beta1_pro <- mean(coverage[2,])
beta2_pro <- mean(coverage[3,])

beta0_pro
```

```
## [1] 0.952
```

```r
beta1_pro
```

```
## [1] 0.942
```

```r
beta2_pro
```

```
## [1] 0.946
```

Comments: Based on the results, we found that average estimates are very close to the true values; the proportion of confidence intervals that cover true value is very close to 0.95 (since we constructed 95% CI).

(d)

```r
N <- 1000
x1 <- rnorm(N, 0, 1)
x2 <- rbinom(N, 1, 0.25)
```

```
estimates <- matrix(NA,nrow=3, ncol=500)
CIs <- matrix(NA,nrow=6,ncol=500)


for (i in 1:500) {

  beta0 <- 0.5
  beta1 <- 0.2
  beta2 <- 0.5
  gamma0 <- 0.5
  gamma1 <- 0
  gamma2 <- 0.5

  y_star <- rpois(N, exp(beta0 + beta1*x1 + beta2*x2))

  pie <- exp(gamma0 + gamma1*x1 + gamma2*x2) /
    (1 + exp(gamma0 + gamma1*x1 + gamma2*x2))

  w <- rbinom(N, 1, pie)
  y <- c(NA, N)
  y[w==1] <- 0
  y[w==0] <- y_star[w==0]


  # fit model and use ad-hoc method
  mod <- glm(y ~ x1 + x2, family = poisson)

  estimate_dispersion <- summary(mod)$deviance / summary(mod)$df.residual

  est_beta0 <- summary(mod)$coefficients[1,1]
  se0_adj <- summary(mod)$coefficients[1,2] * sqrt(estimate_dispersion)

  est_beta1 <- summary(mod)$coefficients[2,1]
  se1_adj <- summary(mod)$coefficients[2,2] * sqrt(estimate_dispersion)

  est_beta2 <- summary(mod)$coefficients[3,1]
  se2_adj <- summary(mod)$coefficients[3,2] * sqrt(estimate_dispersion)

  estimates[1,i] <- est_beta0
  estimates[2,i] <- est_beta1
  estimates[3,i] <- est_beta2

  c <- 1.96
  CIs[1,i] <- est_beta0 - c*se0_adj
  CIs[2,i] <- est_beta0 + c*se0_adj
  CIs[3,i] <- est_beta1 - c*se1_adj
  CIs[4,i] <- est_beta1 + c*se1_adj
  CIs[5,i] <- est_beta2 - c*se2_adj
  CIs[6,i] <- est_beta2 + c*se2_adj

}

# average estimates
```

```r
average_est_beta0 <- mean(estimates[1,])
average_est_beta1 <- mean(estimates[2,])
average_est_beta2 <- mean(estimates[3,])

average_est_beta0
```

```
## [1] -0.4761964
```

```r
average_est_beta1
```

```
## [1] 0.1969907
```

```r
average_est_beta2
```

```
## [1] 0.1531694
```

```r
# proportion of confidence interval coverage
coverage <- matrix(NA, nrow=3, ncol=500)

for (i in 1:500) {
  L0 <- CIs[1,i]
  U0 <- CIs[2,i]
  L1 <- CIs[3,i]
  U1 <- CIs[4,i]
  L2 <- CIs[5,i]
  U2 <- CIs[6,i]

  coverage[1,i] <- (beta0 >= L0) & (U0 >= beta0)
  coverage[2,i] <- (beta1 >= L1) & (U1 >= beta1)
  coverage[3,i] <- (beta2 >= L2) & (U2 >= beta2)
}

beta0_pro <- mean(coverage[1,])
beta1_pro <- mean(coverage[2,])
beta2_pro <- mean(coverage[3,])

beta0_pro
```

```
## [1] 0
```

```r
beta1_pro
```

```
## [1] 0.904
```

```r
beta2_pro
```

```
## [1] 0.218
```

Comments: By using ad-hoc method, it does not yield valid inference. Since the proportion of confidence intervals that cover the true value is till far less than 0.95 after using estimated dispersion parameter.