Assignment 8 - Poisson and Logistic Regression

EPIB 607 - FALL 2021

your name and McGill ID

compiled on December 04, 2021

1 (50 points) Population mortality rates in Denmark

(a)

Determinants
1) Sex: Female (reference), Male
2) Year (time period): 1980-1984 (reference), 2000-2004, 2005-2009
3) Age: 70-74 (reference), 75-79, 80-84, 85-89

$$sex = \begin{cases} 1, & \text{if male} \\ 0, & \text{if female} \end{cases}$$

$$year1 = \begin{cases} 1, & \text{if 2000-2004} \\ 0, & \text{otherwise} \end{cases}$$

$$year2 = \begin{cases} 1, & \text{if 2005-2009} \\ 0, & \text{otherwise} \end{cases}$$

$$age75 = \begin{cases} 1, & \text{if 75-79} \\ 0, & \text{otherwise} \end{cases}$$

$$age80 = \begin{cases} 1, & \text{if 80-84} \\ 0, & \text{otherwise} \end{cases}$$

$$age85 = \begin{cases} 1, & \text{if 85-89} \\ 0, & \text{otherwise} \end{cases}$$

Parameters:
$\boldsymbol{\mu}$: Expected number of deaths in Denmark
$\boldsymbol{\lambda_0}$: Mortality rate for females aged 70-74 during 1980-1984 (reference category) in Denmark
$\boldsymbol{\theta_1}$: Mortality rate ratio for males vs females (reference) in Denmark adjusted for age and year

$\boldsymbol{\theta_2}$: Mortality rate ratio for time period 2000-2004 vs time period 1980-1984 (reference) in Denmark adjusted for age and sex
$\boldsymbol{\theta_3}$: Mortality rate ratio for time period 2005-2009 vs time period 1980-1984 (reference) in Denmark adjusted for age and sex
$\boldsymbol{\theta_4}$: Mortality rate ratio for those aged 75-79 years vs those aged 70-74 years (reference) in Denmark adjusted for year and sex
$\boldsymbol{\theta_5}$: Mortality rate ratio in those aged 80-84 years vs those aged 70-74 years (reference) in Denmark adjusted for year and sex
$\boldsymbol{\theta_6}$: Mortality rate ratio in those aged 85-89 years vs those aged 70-74 years (reference) in Denmark adjusted for year and sex

Regression Model

Since:

$$\mu = \lambda \cdot PT$$

$$\lambda = \lambda_0 \cdot \theta_1^{sex} \cdot \theta_2^{year1} \cdot \theta_3^{year2} \cdot \theta_4^{age75} \cdot \theta_5^{age80} \cdot \theta_6^{age85}$$

Therefore:

$$\mu = \lambda_0 \cdot \theta_1^{sex} \cdot \theta_2^{year1} \cdot \theta_3^{year2} \cdot \theta_4^{age75} \cdot \theta_5^{age80} \cdot \theta_6^{age85} \cdot PT \qquad \text{(multiplicative model)}$$

Regression Equation:

$$\begin{aligned} log(\mu) = log(\lambda_0) + log(\theta_1) \cdot sex + log(\theta_2) \cdot year1 + log(\theta_3) \cdot year2 + \\ log(\theta_4) \cdot age75 + log(\theta_5) \cdot age80 + log(\theta_6) \cdot age85 + log(PT) \end{aligned}$$

(b)

```
##
## Call:
## glm(formula = deaths ~ sex + year1 + year2 + age75 + age80 +
##     age85 + offset(log(pt)), family = poisson(link = "log"),
##     data = data_fit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -14.1385   -4.3278   -0.8631    2.4490   19.1191
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.501419   0.004006 -874.11   <2e-16 ***
## sex          0.420525   0.002887  145.64   <2e-16 ***
## year1       -0.148991   0.003490  -42.70   <2e-16 ***
## year2       -0.252000   0.003557  -70.85   <2e-16 ***
## age75        0.494710   0.004260  116.14   <2e-16 ***
## age80        0.984528   0.004175  235.80   <2e-16 ***
## age85        1.496468   0.004300  348.04   <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: 151058.1  on 23  degrees of freedom
## Residual deviance:   1439.8  on 17  degrees of freedom
## AIC: 1735.3
##
## Number of Fisher Scoring iterations: 3
```

Fitted regression equation

$$\begin{aligned}\widehat{log(\mu)} = -3.5 \ + \ 0.42 \cdot sex \ + \ -0.15 \cdot year1 \ + \ -0.25 \cdot year2 \ + \\ 0.49 \cdot age75 \ + \ 0.98 \cdot age80 \ + \ 1.5 \cdot age85 \ + \ log(PT)\end{aligned}$$

OR

$$\hat{\mu} = 0.03 \cdot 1.52^{sex} \cdot 0.86^{year1} \cdot 0.78^{year2} \cdot 1.64^{age75} \cdot 2.68^{age80} \cdot 4.47^{age85} \cdot PT$$

(c)

The mortality rate ratio in Denmark of males vs females (reference) is 1.52 This means that, on average, the mortality rate among males is 52 % greater than among females while controlling for Age and Time period.
95% CI for gender mortality rate is [1.51, 1.53]
The 95% CI for gender mortality rate doesn't include the null (1) which indicates statistical significance, additionally the 95% CI is narrow and is far away from null (1) suggesting that mortality rates are significantly different in males compared with females in Denmark.

(d)

Null hypothesis for goodness of fit test: there is no lack of fit
Alternative hypothesis for goodness of fit test: there is lack of fit

pvalue for goodness of fit test is $1.2445845 \times 10^{-297}$, so at an $\alpha = 0.05$, we have evidence to reject the null which suggests that there is a lack of fit and that this model is not a good fit.

2 (50 points) Survival of patients following admission to an adult intensive care unit (ICU)

(a)

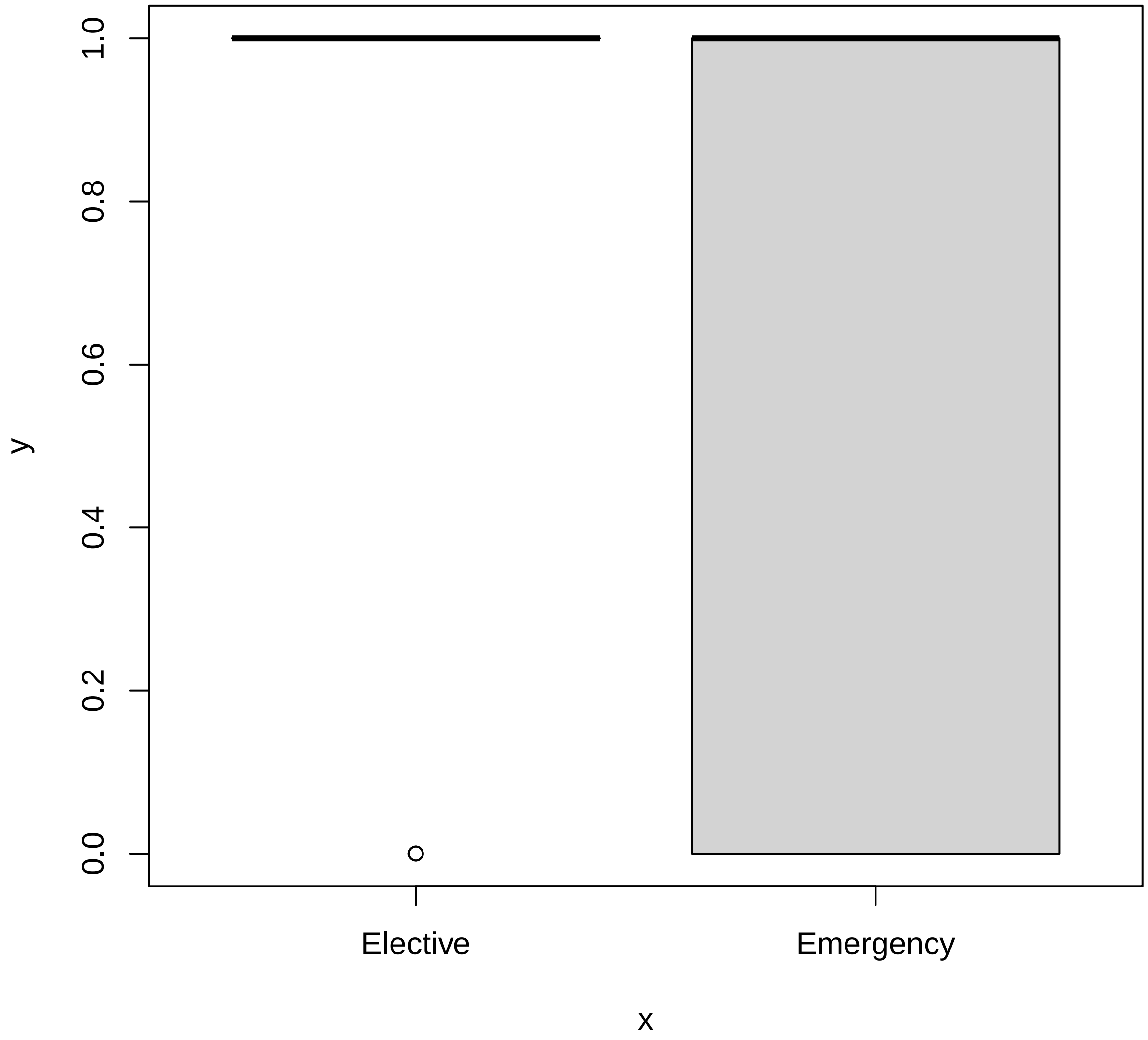
Equation: $survived_i \sim$ Bernoulli(π_i), $logit(\pi_i) = \alpha + \beta * Type_i$, with $Type_i = 0$ if admitted in the elective way, and $Type_i = 1$ if admitted in the emergency way. (4 pts)
(You can calculate that for the probability of death as well)
We use logistic regression because our outcome "sta" is a binary variable. (2 pts)
our parameter of interest is the log odds ratio of emergency admission over elective admission on the discharge status. The odds of death increases from $exp(exp(\alpha))$ to $exp(\alpha + \beta)$ if admitted in emergency. (4 pts)
(The probability of death increased from $logistic(exp(\alpha))$ to $logistic(\alpha + \beta)$ if admitted in emergency.)

(b)

correct plot (2 pts)

```
##  Elective Emergency
##        53       147
```

```
## 160 patients survived, and 40 patients died
```



53 patients were admitted as Elective, 147 were admitted as emergency. (2 pts) Compared to patients admitted as elective, patients admitted as emergency are less likely to survive. (96.2% elective patients survived while only 74.1% emergency patients survived) (4 pts)

(c)

```
##
## Call:
## glm(formula = sta ~ type, family = binomial, data = icu)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7734  -0.7734  -0.7734  -0.2774   2.5601
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.2387     0.7206  -4.495 6.97e-06 ***
## typeEmergency   2.1849     0.7448   2.934  0.00335 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 185.05  on 198  degrees of freedom
## AIC: 189.05
##
## Number of Fisher Scoring iterations: 5
```

```
##                    2.5 %    97.5 %
## (Intercept)   -5.0492011 -2.070861
## typeEmergency  0.9486731  4.024670
```

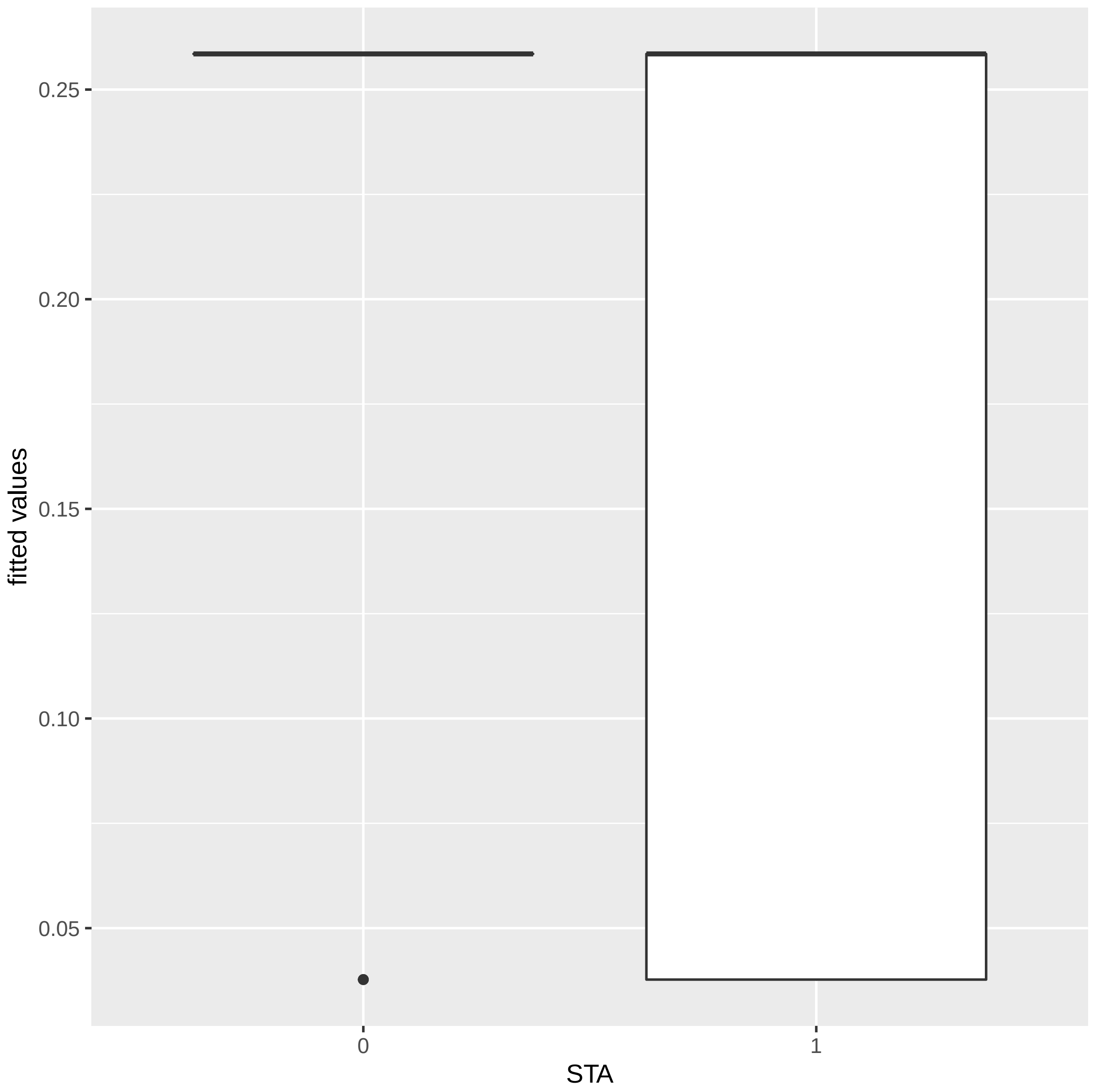
Assumptions: 1. independence between observations, patients in the same admission group have equal probability of survival (just like in binomial distribution) (2 pts) 2. sample size large enough (2 pts) the log-odds ratio is $\beta = -2.1849$, 95% CI for β is (-4.02, -0.95). (4 pts) Based on $\alpha = 0.05$, we reject the null hypothesis: the odds of survival given emergency is less than that given elective (p-value = 0.00335). (2 pts)

(d)

$pi_i = exp(\alpha + \beta * Type_i)/(1 + exp(\alpha + \beta * Type_i))$ (4 pts) Elective patient: exp(3.2387)/(1+ exp(3.2387)) = 0.962; Emergency patient: exp(3.2387 - 2.1849)/(1+ exp(3.2387 - 2.1849)) = 0.741 (4 pts) The estimated survival probability given elective is 0.962, that gicen emergency is 0.741. (2 pts)

(e)

correct plot (2 pts)



We can see that most of the fitted survival probability for the survived patients are close to 1, while that for the other patients are smaller. (2 pts) Thus, our fitted values matches the truth. We have a very wierd boxplot because we only have 2 fitted values. Since we only include one independent variable “type” and given “type” also binary, there are only 2 possible fitted values available. (2 pts)

(f)

```
##
## Call:
```

```
## glm(formula = type ~ sta, family = binomial, data = icu)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4478  -1.5122   0.8762   0.8762   0.8762
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.7595     0.1697   4.477 7.57e-06 ***
## staDied       2.1849     0.7450   2.933  0.00336 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 231.29  on 199  degrees of freedom
## Residual deviance: 216.18  on 198  degrees of freedom
## AIC: 220.18
##
## Number of Fisher Scoring iterations: 5
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.4329134 1.099546
## staDied     0.9486752 4.024670
```

difference: we switched the position of dependent and independent variable within the logistic regression model. (2 pts)
similarity: the value of the estimates (α, β) remains the same. (2 pts)

I prefer the first model, because death is obviously a outcome depending on patient's previous health condition (admission type). (2 pts)

Code

```
## ---- Setup -------------------------------------------------------------------
# set default chunk options here
knitr::opts_chunk$set(
  echo = FALSE,          # don’t show code
  warning = FALSE,       # don’t show warnings
  message = FALSE,       # don’t show messages (less serious warnings)
  cache = TRUE,         # set to TRUE to save results from last compilation
  fig.align = "center",   # center figures
  fig.asp = 1          # fig.aspect ratio
)
library(tidyverse)
## ---- Question-1 -------------------------------------------------------------
denmark <- read.csv("denmark.csv")

data_fit <- denmark %>%
  select(-contains(c("Male", "rate"), ignore.case = F)) %>%
  transmute(year1 = ifelse(Year %in% "2000-2004", 1, 0),
            year2 = ifelse(Year %in% "2005-2009", 1, 0),
            age75 = ifelse(Age %in% "75-79", 1, 0),
            age80 = ifelse(Age %in% "80-84", 1, 0),
            age85 = ifelse(Age %in% "85-89", 1, 0),
            sex = 0,
            pt = Female_PT,
            deaths = Female_deaths) %>%
  rbind.data.frame(denmark %>%
  select(-contains(c("Female", "rate"), ignore.case = F)) %>%
  transmute(year1 = ifelse(Year %in% "2000-2004", 1, 0),
            year2 = ifelse(Year %in% "2005-2009", 1, 0),
            age75 = ifelse(Age %in% "75-79", 1, 0),
            age80 = ifelse(Age %in% "80-84", 1, 0),
            age85 = ifelse(Age %in% "85-89", 1, 0),
            sex = 1,
            pt = Male_PT,
            deaths = Male_deaths))

fit1 <- glm(deaths ~ sex + year1 + year2 + age75 + age80 + age85 + offset(log(pt)),
            data = data_fit, family = poisson(link = "log"))
sum_fit1 <- summary(fit1); sum_fit1
beta0 <- round(coefficients(sum_fit1)[1, 1], 2)
beta1 <- round(coefficients(sum_fit1)[2, 1], 2)
beta2 <- round(coefficients(sum_fit1)[3, 1], 2)
beta3 <- round(coefficients(sum_fit1)[4, 1], 2)
```

```r
beta4 <- round(coefficients(sum_fit1)[5, 1], 2)
beta5 <- round(coefficients(sum_fit1)[6, 1], 2)
beta6 <- round(coefficients(sum_fit1)[7, 1], 2)
lambda0 <- round(exp(coefficients(sum_fit1)[1, 1]), 2)
theta1 <- round(exp(coefficients(sum_fit1)[2, 1]), 2)
theta2 <- round(exp(coefficients(sum_fit1)[3, 1]), 2)
theta3 <- round(exp(coefficients(sum_fit1)[4, 1]), 2)
theta4 <- round(exp(coefficients(sum_fit1)[5, 1]), 2)
theta5 <- round(exp(coefficients(sum_fit1)[6, 1]), 2)
theta6 <- round(exp(coefficients(sum_fit1)[7, 1]), 2)

ci_theta1 <- round(exp(confint(fit1)[2, ]), 2)
# 3 ways to calculate fitted values
# (all valid and all will result in the same answer)
data_fit$expected <- fitted(fit1)
#or by hand using multiplicative model
exp_deaths <- (lambda0) * (theta1^data_fit$sex) * (theta2^data_fit$year1) *
  (theta3^data_fit$year2) * (theta4^data_fit$age75) * (theta5^data_fit$age80) *
  (theta6^data_fit$age85) * data_fit$pt
#or by hand using regression equation (model on log scale)
exp_deaths <- exp(beta0 + beta1*data_fit$sex + beta2*data_fit$year1 +
                    beta3*data_fit$year2 + beta4*data_fit$age75 +
                    beta5*data_fit$age80 + beta6*data_fit$age85 +
                    log(data_fit$pt))

chi_stat <- sum(((data_fit$deaths - data_fit$expected)^2) / data_fit$expected)
p_value_fit <- pchisq(q = chi_stat,
                      df = nrow(data_fit) - length(coef(fit1)),
                      lower.tail = F)
## ---- Question-2 -------------------------------------------------------------
load("icu.rda")

icu$sta <- ifelse(icu$sta == "Lived",1,0)
summary(icu$type)
cat(sum(icu$sta),"patients survived, and",(length(icu$sta)-sum(icu$sta)),"patients died'
plot(icu$type,icu$sta)
model1 <- glm(sta ~ type, data = icu, family = binomial)
summary(model1)
confint(model1)
library(ggplot2)
temp_df <- data.frame(cbind(icu$sta,model1$fitted.values))
ggplot(data = temp_df) + geom_boxplot(aes(x=as.factor(X1), y = X2)) + xlab("STA") + yla
model2 <- glm(type ~ sta, data = icu, family = binomial)
summary(model2)
```

```r
confint(model2)
```