Homework 4 Answers

BSTA 513/613

Nicky Wakim

Questions Part 1

Question 1

In this problem, we will practice performing model diagnostics in a logistic regression model. You will need to download and source the `Logistic_Dx_Functions.R` *file from the shared Data folder.*

This question is taken from the Hosmer and Lemeshow textbook. The ICU study data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The dataset should be available in our shared folder. The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. In this question, the primary outcome variable is vital (survival) status at hospital discharge, STA. Clinicians associated with the study felt that a key determinant of survival was the patient's age at admission, AGE. We will build to a multivariable logistic regression model while adjusting for cancer part of the present problem (CAN), CPR prior to ICU admission (CPR), infection probable at ICU admission (INF), and level of consciousness at ICU admission (LOC).

A code sheet for the variables to be considered is displayed in Table 1.5 below (from the Hosmer and Lemeshow textbook, pg. 23). We refer to this data set as the ICU data.

You will need to use some of the mutations implemented in HW 2, Q2, Part d.

Table 1.5 Code Sheet for the Variables in the ICU Data

Variable	Description	Codes/Values	Name
1	Identification code	ID number	ID
2	Vital status at hospital discharge	0 = Lived 1 = Died	STA
3	Age	Years	AGE
4	Gender	0 = Male 1 = Female	GENDER
5	Race	1 = White 2 = Black 3 = Other	RACE
6	Service at ICU admission	0 = Medical 1 = Surgical	SER
7	Cancer part of present problem	0 = No 1 = Yes	CAN
8	History of chronic renal failure	0 = No 1 = Yes	CRN
9	Infection probable at ICU admission	0 = No 1 = Yes	INF
10	CPR prior to ICU admission	0 = No 1 = Yes	CPR
11	Systolic blood pressure at ICU admission	mm Hg	SYS
12	Heart rate at ICU admission	Beats/min	HRA
13	Previous admission to an ICU within 6 months	0 = No 1 = Yes	PRE
14	Type of admission	0 = Elective 1 = Emergency	TYPE
15	Long bone, multiple, neck, single area, or hip fracture	0 = No 1 = Yes	FRA
16	PO_2 from initial blood gases	0 = >60 1 = ≤60	PO2
17	PH from initial blood gases	0 = ≥7.25 1 = <7.25	PH
18	PCO_2 from initial blood gases	0 = ≤45 1 = >45	PCO
19	Bicarbonate from initial blood gases	0 = ≥18 1 = <18	BIC
20	Creatinine from initial blood gases	0 = ≤2.0 1 = >2.0	CRE
21	Level of consciousness at ICU admission	0 = No coma or deep stupor 1 = Deep stupor 2 = Coma	LOC

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-1.924	0.284	-6.764	0.000	-2.521	-1.399
CANYes	0.093	0.614	0.151	0.880	-1.246	1.219
CPRYes	1.514	0.602	2.516	0.012	0.326	2.730
INFYes	0.807	0.371	2.176	0.030	0.084	1.547

We will use the following model:

$$\text{logit}(\pi(\mathbf{X})) = \beta_0 + \beta_1 \cdot I(CAN = \text{“Yes”}) + \beta_2 \cdot I(CPR = \text{“Yes”}) + \beta_3 \cdot I(INF = \text{“Yes”})$$

```
icu = read_csv(here("data", "icu.csv"))
icu1 = icu %>% mutate(STA = as.factor(STA) %>% relevel(ref = "Lived"))
icu2 = icu1 %>% mutate(CAN = as.factor(CAN) %>% relevel(ref = "No"),
                       CPR = as.factor(CPR) %>% relevel(ref = "No"),
                       INF = as.factor(INF) %>% relevel(ref = "No"),
                       LOC = as.factor(LOC) %>%
                         relevel(ref = "No Coma or Deep Stupor"))
```

I will fit the regression equation here.

```
model1 = glm(STA ~ CAN + CPR + INF,
             data = icu2, family = binomial)
m1_tidy = tidy(model1, conf.int = T)

m1_tidy %>% gt() %>%
  tab_options(table.font.size = 16) %>%
  fmt_number(decimals = 3)
```

Now I can use it throughout this question.

I will also load the R script needed to use the `dx()` function.

```
source(here("data", "Logistic_Dx_Functions.R"))
```

I can run the model through the `dx()` function:

```
dx_icu = dx(model1)
glimpse(dx_icu)
```

```
Rows: 7
Columns: 16
$ `(Intercept)` <dbl> 1, 1, 1, 1, 1, 1, 1
$ CANYes        <dbl> 0, 0, 1, 0, 1, 0, 1
$ CPRYes        <dbl> 0, 0, 0, 1, 0, 1, 1
$ INFYes        <dbl> 1, 0, 1, 0, 0, 1, 1
$ y             <dbl> 17, 12, 1, 1, 3, 6, 0
$ P             <dbl> 0.2465297, 0.1273696, 0.2641879, 0.3989183, 0.1380566, 0~
$ n             <int> 69, 99, 6, 4, 13, 8, 1
$ yhat          <dbl> 17.0105460, 12.6095911, 1.5851272, 1.5956732, 1.7947357,~
$ Pr            <dbl> -0.002945748, -0.183769303, -0.541794689, -0.608232077, ~
$ dr            <dbl> -0.002945953, -0.185061578, -0.568559201, -0.627265966, ~
$ h             <dbl> 0.013276657, 0.008995739, 0.074100939, 0.093495654, 0.04~
$ sPr           <dbl> -0.0029655, -0.1846015, -0.5630577, -0.6388286, 0.991157~
$ sdr           <dbl> -0.002965706, -0.185899619, -0.590872626, -0.658820007, ~
$ dChisq        <dbl> 8.794188e-06, 3.407771e-02, 3.170340e-01, 4.081020e-01, ~
$ dDev          <dbl> 8.795411e-06, 3.455867e-02, 3.491305e-01, 4.340438e-01, ~
$ dBhat         <dbl> 1.183284e-07, 3.093369e-04, 2.537265e-02, 4.209110e-02, ~
```

Part a

How many covariate patterns does the regression equation have?

There are three different covariates in the regression equation: CAN, CPR, and INF. Each of the three covariates have 2 categories.

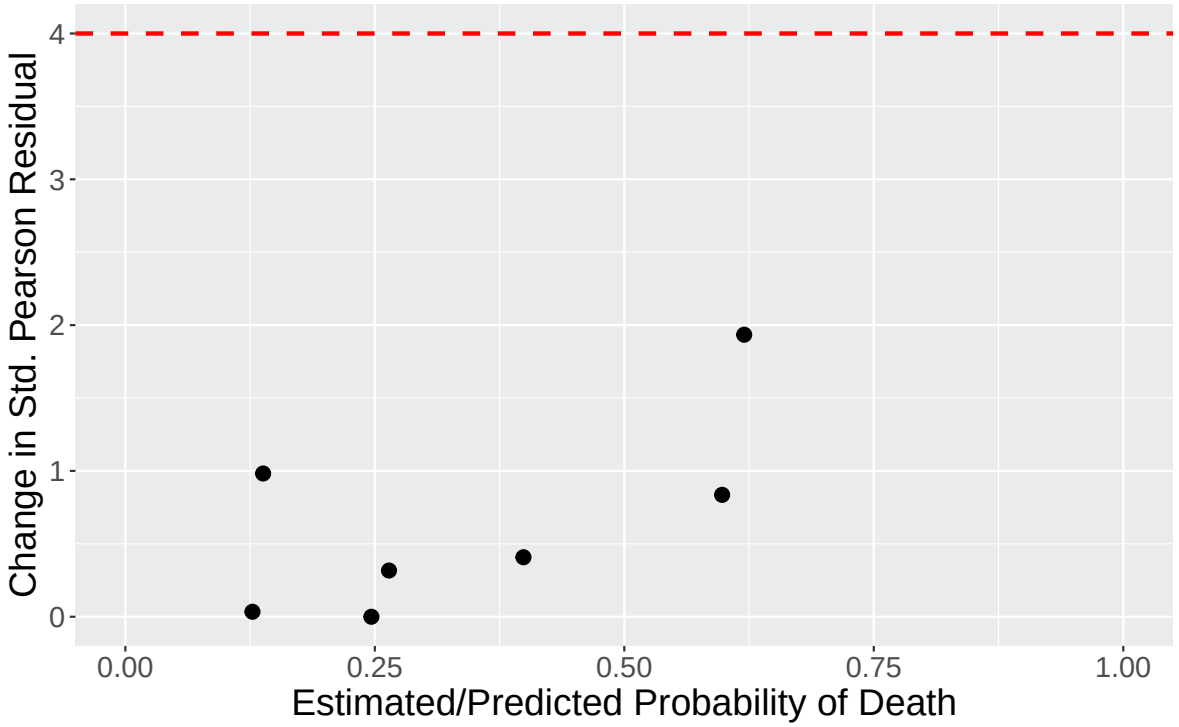
There are 8 possible combinations of these 3 variables with 2 categories. If you are familiar with combinatorics, this is 3 choose 2. If you are not familiar with combinatorics, then we can think about a table with possible combinations:

Covariate Pattern	CAN	CPR	INF
1	Yes	Yes	Yes
2	Yes	Yes	No
3	Yes	No	Yes
4	Yes	No	No
5	No	Yes	Yes
6	No	Yes	No
7	No	No	Yes
8	No	No	No

Part b

Plot the change in standardized Pearson residual by predicted probability. Do you notice any potential outliers? Explain your reasoning.

```
ggplot(dx_icu) +
  geom_hline(yintercept = 4, col = "red", linetype="dashed", linewidth=1) +
  geom_point(aes(x=P, y=dChisq), size = 3) +
  xlab("Estimated/Predicted Probability of Death") +
  ylab("Change in Std. Pearson Residual") +
  theme(text = element_text(size = 18)) +
  xlim(0,1)
```

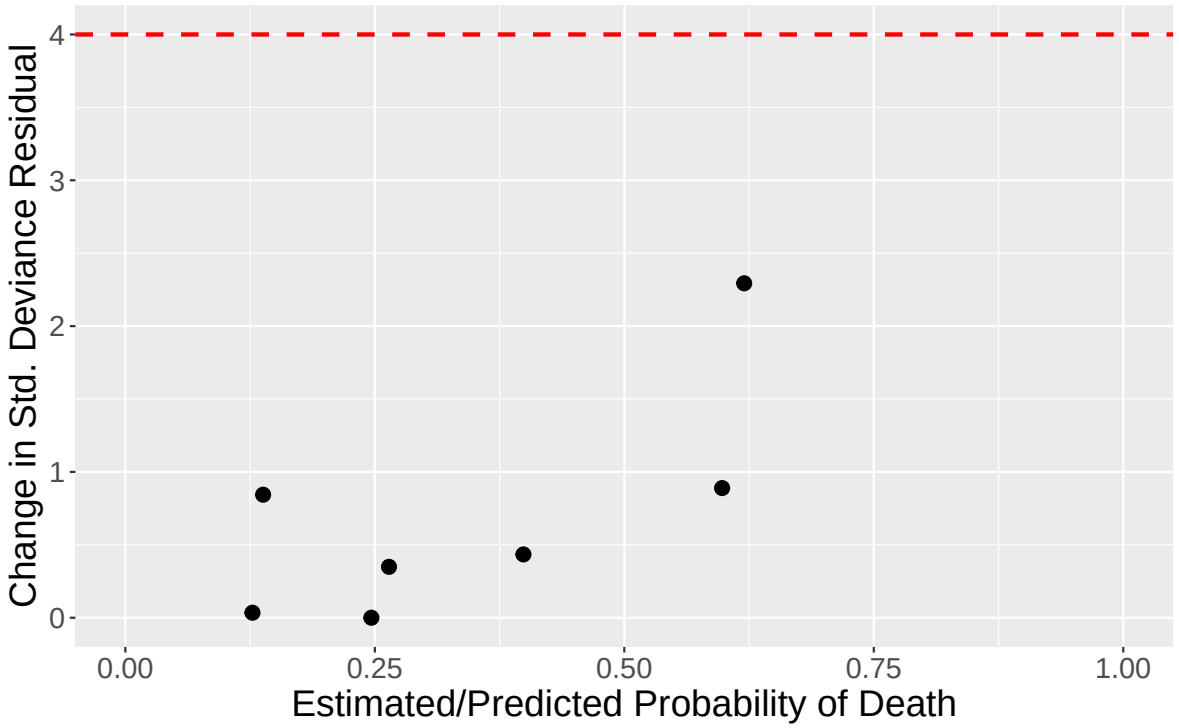


Part c

Plot the change in standardized Deviance residual by predicted probability. Do you notice any potential outliers? Explain your reasoning.

```
ggplot(dx_icu) +
  geom_hline(yintercept = 4, col = "red", linetype="dashed", linewidth=1) +
  geom_point(aes(x=P, y=dDev), size = 3) +
  xlab("Estimated/Predicted Probability of Death") +
```

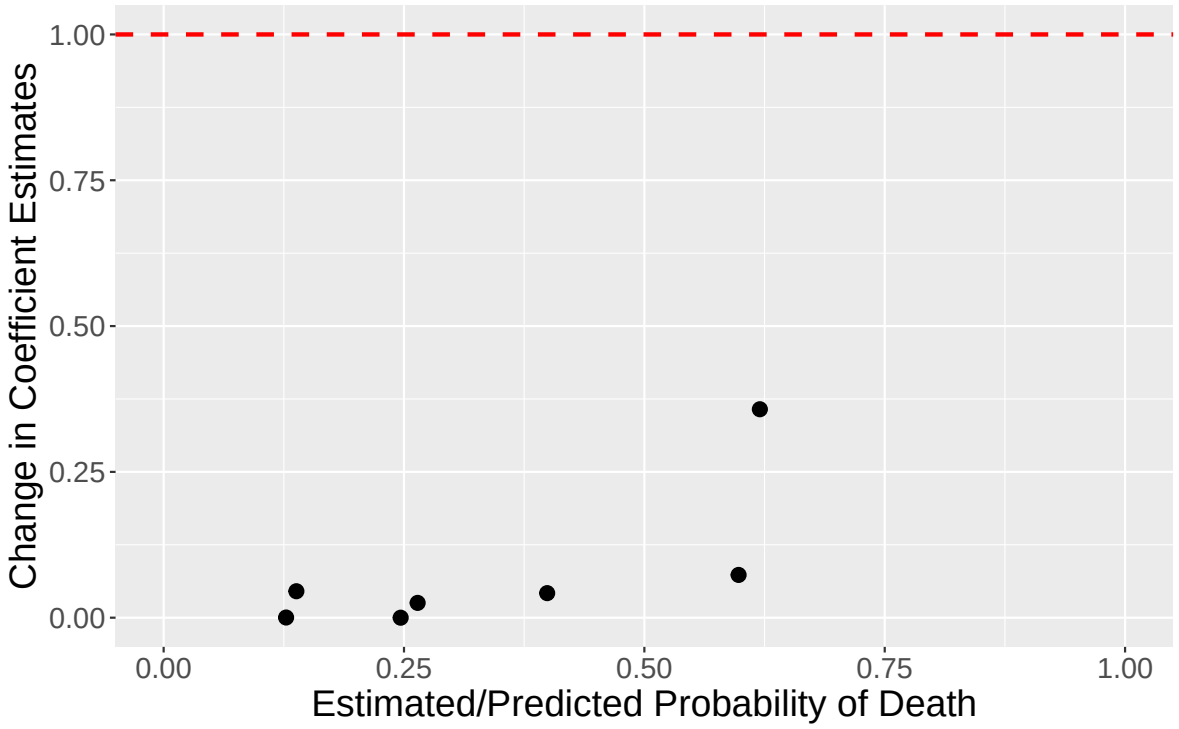
```
  ylab("Change in Std. Deviance Residual") +
  theme(text = element_text(size = 18)) + xlim(0, 1)
```



Part d

Plot the change in coefficient estimate by predicted probability. Do you notice any influential points? Explain your reasoning.

```
ggplot(dx_icu) +
  geom_hline(yintercept = 1, col = "red", linetype="dashed", linewidth=1) +
  geom_point(aes(x=P, y=dBhat), size = 3) +
  xlab("Estimated/Predicted Probability of Death") +
  ylab("Change in Coefficient Estimates") +
  theme(text = element_text(size = 18)) + xlim(0,1)
```

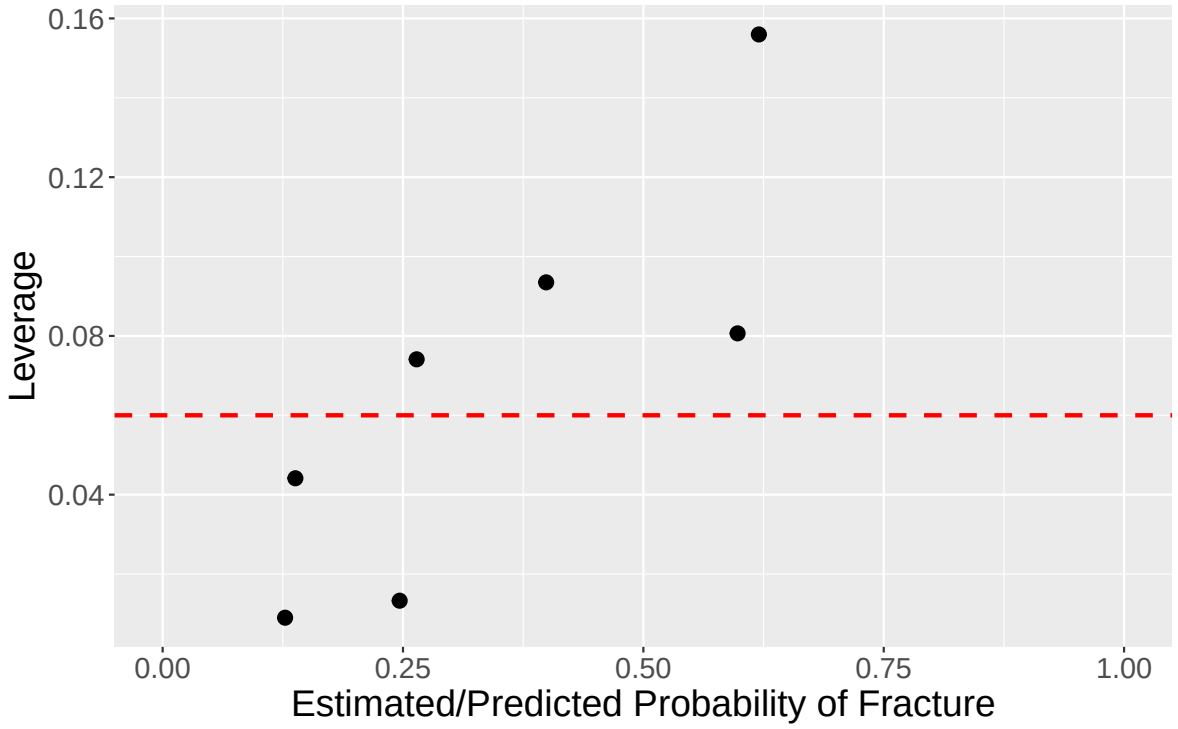


Part e

Plot the leverage by predicted probability. Do you notice any influential points? Explain your reasoning.

```
n = nrow(icu2)

ggplot(dx_icu) +
  geom_hline(yintercept = 3*4/n, col = "red", linetype="dashed", linewidth=1) +
  geom_point(aes(x=P, y=h), size=3) +
  xlab("Estimated/Predicted Probability of Fracture") +
  ylab("Leverage") +
  theme(text = element_text(size = 18)) + xlim(0, 1)
```



Question 2

In this problem, we will practice fitting and interpreting a ***log-binomial regression.***

This question is taken from the Hosmer and Lemeshow textbook. The ICU study data set consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The dataset should be available in our shared folder. The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. In this question, the primary outcome variable is vital (survival) status at hospital discharge, STA. Clinicians associated with the study felt that a key determinant of survival was the patient's age at admission, AGE. We will build to a multivariable logistic regression model while adjusting for cancer part of the present problem (CAN), CPR prior to ICU admission (CPR), infection probable at ICU admission (INF), and level of consciousness at ICU admission (LOC).

A code sheet for the variables to be considered is displayed in Table 1.5 below (from the Hosmer and Lemeshow textbook, pg. 23). We refer to this data set as the ICU data.

You will need to use some of the mutations implemented in HW 2, Q2, Part d.

Table 1.5 Code Sheet for the Variables in the ICU Data

Variable	Description	Codes/Values	Name
1	Identification code	ID number	ID
2	Vital status at hospital discharge	0 = Lived 1 = Died	STA
3	Age	Years	AGE
4	Gender	0 = Male 1 = Female	GENDER
5	Race	1 = White 2 = Black 3 = Other	RACE
6	Service at ICU admission	0 = Medical 1 = Surgical	SER
7	Cancer part of present problem	0 = No 1 = Yes	CAN
8	History of chronic renal failure	0 = No 1 = Yes	CRN
9	Infection probable at ICU admission	0 = No 1 = Yes	INF
10	CPR prior to ICU admission	0 = No 1 = Yes	CPR
11	Systolic blood pressure at ICU admission	mm Hg	SYS
12	Heart rate at ICU admission	Beats/min	HRA
13	Previous admission to an ICU within 6 months	0 = No 1 = Yes	PRE
14	Type of admission	0 = Elective 1 = Emergency	TYPE
15	Long bone, multiple, neck, single area, or hip fracture	0 = No 1 = Yes	FRA
16	PO_2 from initial blood gases	0 = >60 1 = ≤60	PO2
17	PH from initial blood gases	0 = ≥7.25 1 = <7.25	PH
18	PCO_2 from initial blood gases	0 = ≤45 1 = >45	PCO
19	Bicarbonate from initial blood gases	0 = ≥18 1 = <18	BIC
20	Creatinine from initial blood gases	0 = ≤2.0 1 = >2.0	CRE
21	Level of consciousness at ICU admission	0 = No coma or deep stupor 1 = Deep stupor 2 = Coma	LOC

Part a

Write down the population equation for the log-binomial regression model of STA on AGE, CAN, CPR, and INF. How many parameters does this model contain?

The main thing to remember that we need to use the *log* instead of the *logit*

$$\log(\pi(\mathbf{X})) = \beta_0 + \beta_1 \cdot AGE + \beta_2 \cdot I(CAN = \text{“Yes”}) + \beta_3 \cdot I(CPR = \text{“Yes”}) + \beta_4 \cdot I(INF = \text{“Yes”})$$

This model contains 4 parameters.

Part b

Try using `glm()` to obtain the maximum likelihood estimates of the parameters of the log-binomial regression model in Part a. Do you run into any issues using `glm()`? If you try `logbin()`, does it fix the issues? Explain why and what warnings `logbin()` gives you.

Hint: keep the `glm()` function in its own code chunk so you can add `#| eval: false`. `glm()` may throw an error, so we want to show the work for `glm()` even though it'll break your qmd rendering.

```
logbin_mod <- glm(STA ~ AGE + CAN + CPR + INF,
                  data = icu2, family = binomial(link="log"))
```

`glm` does not work. I get an error `Error: no valid set of coefficients has been found: please supply starting values`. We could try to give it starting values, but I think `logbin()` is a better try.

```
library(logbin)
logbin_mod1 <- logbin(STA ~ AGE + CAN + CPR + INF,
                  data = icu2)
```

```
Warning: nplbin: algorithm did not converge within 10000 iterations -- increase
'maxit'.
```

```
Warning: nplbin: fitted probabilities numerically 1 occurred
```

```
summary(logbin_mod1)
```

```
Warning: MLE on boundary of parameter space, cannot use asymptotic covariance
matrix
```

```
Call:
logbin(formula = STA ~ AGE + CAN + CPR + INF, data = icu2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3088  -0.6806  -0.5552  -0.3735   2.3302

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.043e+00        NaN     NaN      NaN
AGE          1.727e-02        NaN     NaN      NaN
CANYes      -2.187e-11        NaN     NaN      NaN
CPRYes       9.241e-01        NaN     NaN      NaN
INFYes       5.297e-01        NaN     NaN      NaN

    Null deviance: 200.16  on 199  degrees of freedom
Residual deviance: 179.24  on 195  degrees of freedom

  AIC: 189.24
AIC_c: 189.55

Number of iterations: 60845 (best: 827)
```

I also get warnings here. The warning is telling me that the model fitting did not converge. This means the function did not necessarily find the best fit.

Part c

Using `logbin()`, obtain the maximum likelihood estimates of the parameters of the log-binomial regression model in Part a, **but now take out age**. Using these estimates, write down the equation with the fitted values.

```r
logbin_mod2 <- logbin(STA ~ CAN + CPR + INF,
                      data = icu2)
summary(logbin_mod2)
```

Part d

Interpret the exponential of the coefficient (risk ratio) estimate for `CPR`.

Part e

We were not able to fit a model with age, but let's just entertain a scenario here. Let's say we fit the model in Part a and got a coefficient estimate of 0.29 with a 95% confidence interval of 0.23 to 0.35. Using the model is Part a, interpret the exponential of the coefficient (risk ratio) estimate for `AGE`.

Questions Part 2

Question 3

For each of the following outcomes, what type of regression would you use? Explain your answer.

Part a

Number of minutes of moderate-to-vigorous physical activity per week (range: 0 to 600+)

Which regression model is most appropriate?

a. Linear regression
b. Logistic regression
c. Log-binomial regression
d. Poisson regression
e. Multinomial logistic regression

a. Linear regression

Part b

Whether the participant meets CDC recommendations for weekly physical activity (Yes/No)

Which regression model is most appropriate?

a. Linear regression
b. Logistic regression
c. Log-binomial regression
d. Poisson regression
e. Multinomial logistic regression

b. Logistic regression
c. Log-binomial regression

Part c

Number of workouts the person completed in the past week (values: 0, 1, 2, …, 14)

Which regression model is most appropriate?

a. Linear regression
b. Logistic regression
c. Log-binomial regression
d. Poisson regression
e. Multinomial logistic regression

d. Poisson regression

Part d

Self-reported activity level: Sedentary, Moderately active, Highly active

Which regression model is most appropriate?

a. Linear regression
b. Logistic regression
c. Log-binomial regression
d. Poisson regression
e. Multinomial logistic regression

e. Multinomial logistic regression