EX9

roi hezkiyahu

6 5 2022

```
# imports
library(tidyverse)
library(glue)
library(tidymodels)
```

Q1

שאלה 1

נתון מודל רגרסיה לוגיסטית עם וקטור פרמטרים β בגודל K, ולו אומד נראות מקסימלית $\hat{\beta}$.

בנוסף, יהי $\hat{\tilde{\beta}}$ תת וקטור של $\hat{\beta}$ בגודל $K_1 < K$.

א. הראו כי $\hat{\tilde{\beta}} \sim N_{K_1}(\tilde{\beta}, \tilde{\Sigma})$ בקירוב, כאשר $\tilde{\Sigma}$ זוהי תת מטריצה של Σ בגודל $K_1 \times K_1$ אשר מכילה את השורות והעמודות של Σ שמתאימות ל $\tilde{\beta}$.

ב. הראו כי $\tilde{\Sigma}^{-1/2}(\hat{\tilde{\beta}} - \tilde{\beta}) \sim N_{K_1}(0, I)$.

רמזים:

- אם $U \sim N_K(\mu, \Sigma_U)$ אז לכל מטריצה קבועה $A_{K_1 \times K}$ מתקיים $AU \sim N_{K_1}(A\mu, A\Sigma_U A^T)$.
- ייתכן ותרצו להשתמש בתכונה המוכרת: $\Sigma^{-\frac{1}{2}}\Sigma^{-\frac{1}{2}} = \Sigma^{-1}$
- ייתכן ותרצו להשתמש בתכונה המוכרת: $\Sigma\Sigma^{-1} = I$

a

$$\text{w.l.o.g assume } \tilde{\beta} = \beta_{1:K_1}$$
$$\text{thus we get: } E(\hat{\tilde{\beta}}) = E(\hat{\beta}_{1:K_1}) = \beta_{1:K_1} = \tilde{\beta}$$
$$\tilde{\Sigma}_{ij} = Cov(\hat{\tilde{\beta}}_i, \hat{\tilde{\beta}}_j) = Cov(\hat{\beta}_i, \hat{\beta}_j) = \Sigma_{ij}$$
$$\text{thus } \tilde{\Sigma} \text{ is a sub matrix of } \Sigma \text{ with rows and columns of } \Sigma \text{ corrosponing to } \tilde{\beta}$$

b

$$\hat{\tilde{\beta}} \sim N(\tilde{\beta}, \tilde{\Sigma})$$
$$\text{from linearity of expected value we get: } \hat{\tilde{\beta}} - \tilde{\beta} \sim N(0, \tilde{\Sigma})$$
$$\text{from the properties of variance and matrix multiplications we get: } V(\tilde{\Sigma}^{-1/2}(\hat{\tilde{\beta}} - \tilde{\beta})) = \tilde{\Sigma}^{-1/2}\tilde{\Sigma}(\tilde{\Sigma}^{-1/2})^t = \tilde{\Sigma}^{-1/2}\tilde{\Sigma}\tilde{\Sigma}^{-1/2} = I$$
$$\text{thus } \tilde{\Sigma}^{-1/2}(\hat{\tilde{\beta}} - \tilde{\beta}) \sim N(0, I)$$

Q2

שאלה 2

בשאלה זו תבצעו סימולציות על מנת לבחון את ההתפלגות האסימפטוטית של האומד למקדם במודל רגרסיה לוגיסטית.

עליכם לבצע את הסימולציה המתוארת מטה שלוש פעמים, בכל פעם עבור גודל מדגם שונה.

א. בצעו את הסימולציה עם $n = 100$.

ב. בצעו את הסימולציה עם $n = 1,000$.

ג. בצעו את הסימולציה עם $n = 10,000$.

ד. השוו בין שלושת הסימולציות:

I. האם הממוצעים, סטיות התקן וההיסטוגרמה השתנו כאשר גודל המדגם גדל? הסבירו.
II. כיצד באה לביטוי בתוצאות הסימולציות העובדה ש $\hat{\beta}_1$ הינו אומד עקיב? הסבירו.
III. האם נראה סביר ש $\hat{\beta}_1$ מתפלג אסימפטוטית נורמלית?

הסימולציה:
בצעו תהליך של יצירת מדגם 1,000 פעמים. בכל איטרציה צרו מדגם בגודל n עם משתנה מסביר אחד (x_1) ומשתנה תלוי (y). התאימו מודל רגרסיה לוגיסטית ושימרו את האומד למקדם המתקבל ואת סטיית התקן שלו. כך תעשו זאת:

- דגמו וקטור x_1 בגודל n של משתנים נורמליים סטנדרטיים $x_1 \sim N(0,1)$.
- צרו וקטור סיכויים p בגודל n על ידי $p = \text{expit}(\beta_0 + \beta_1 x_1)$, עם $\beta_0 = 0.5$ ו- $\beta_1 = 2$.
- צרו וקטור y בגודל n שנדגם מהתפלגות ברנולי עם p.
- התאימו מודל רגרסיה לוגיסטית לנתונים שיצרתם: $y \sim x_1$.
- חלצו את $\hat{\beta}_1$ (האומד למקדם של x_1), ואת סטיית התקן שלו ממודל הרגרסיה ושימרו אותם.

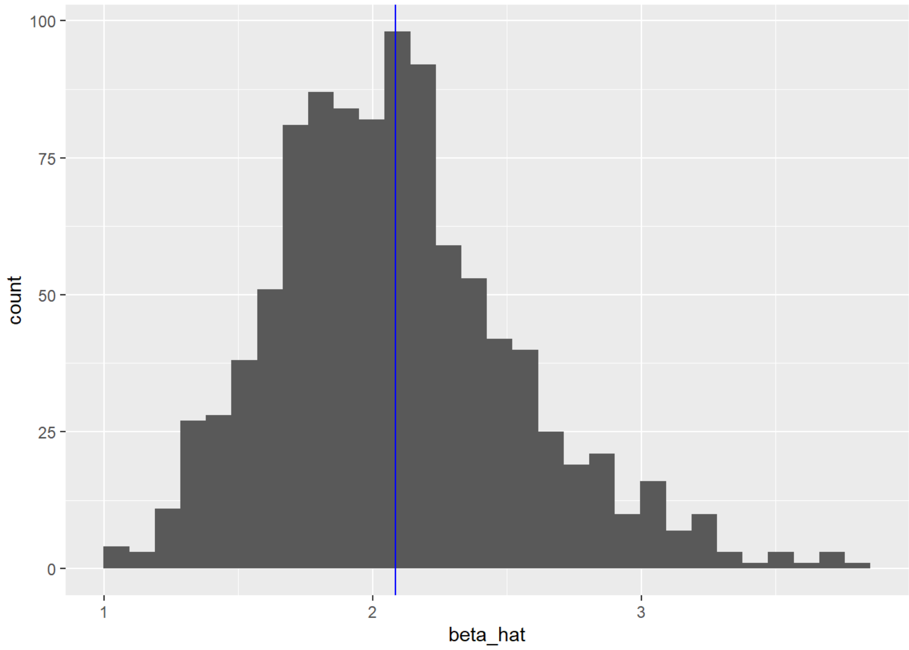
מכל אחד מ1000 האיטרציות של הסימולציה קיבלתם $\hat{\beta}_1$ ואת האומד לסטיית התקן שלו. נרצה לבחון אותם -

- דווחו את הממוצע של ערכי האומדים $\hat{\beta}_1$, ואת הממוצע של האומדנים לסטיית התקן של $\hat{\beta}_1$, שקיבלתם על פני 1000 האיטרציות.
- השוו את ממוצע אומדני סטיית התקן, לסטיית התקן של האומדנים.
- צרו היסטוגרמה של $\hat{\beta}_1$.

```
expit <- function(p){exp(p)/(1+exp(p))}

sim <- function(n){
  beta_hats = c()
  beta_stds = c()
  for (i in 1:1000){
    x1 <- rnorm(n)
    p <- expit(0.5 + 2*x1)
    y <- rbinom(n,1,p)
    model<- glm(y~x1,family = "binomial")
    coef_mat <- tidy(model)
    beta_hat <- coef_mat$estimate[2]
    beta_std <- coef_mat$std.error[2]
    beta_hats[i] <- beta_hat
    beta_stds[i] <- beta_std
  }
  sigma = sqrt(solve(t(x1)%*%diag((p*(1-p)))%*%x1)[1,1])

  return(list(beta_hats,beta_std,sigma))
}

for (n in c(100,1000,10000)){
  res <- sim(n)
  print(glue("the mean for beta for {n} observations is: {mean(res[[1]])}
       the mean for beta std for {n} observations is: {mean(res[[2]])}
       the real std is: {res[[3]]}"))

  print(tibble(beta_hat = res[[1]]) %>%
    ggplot(aes(x = beta_hat))+
    geom_histogram()+
    geom_vline(xintercept =  mean(res[[1]]),color = "blue"))
}
```

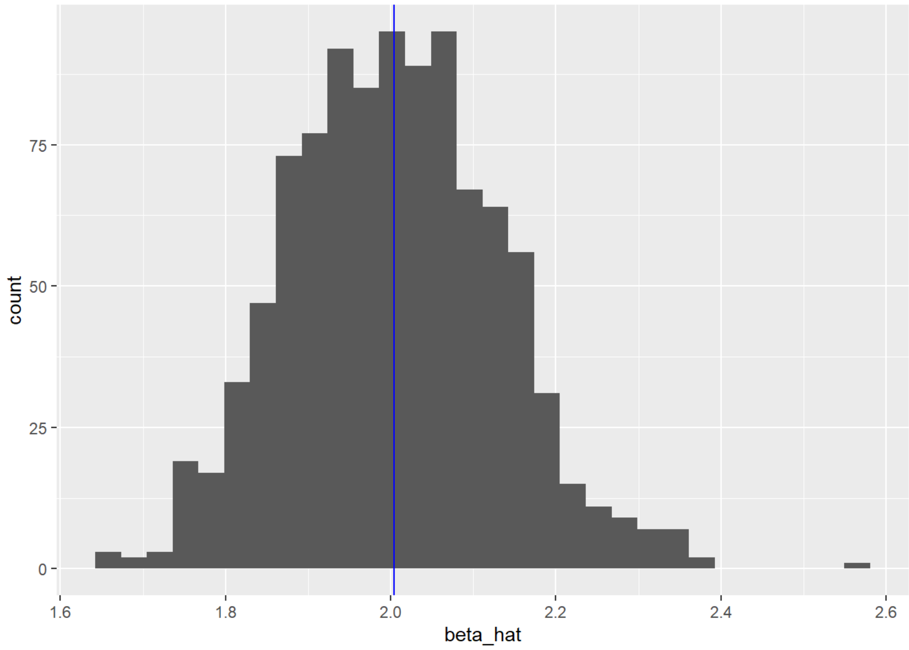
```
## the mean for beta for 100 observations is: 2.08401627076864
## the mean for beta std for 100 observations is: 0.539465146514572
## the real std is: 0.384243969528711
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



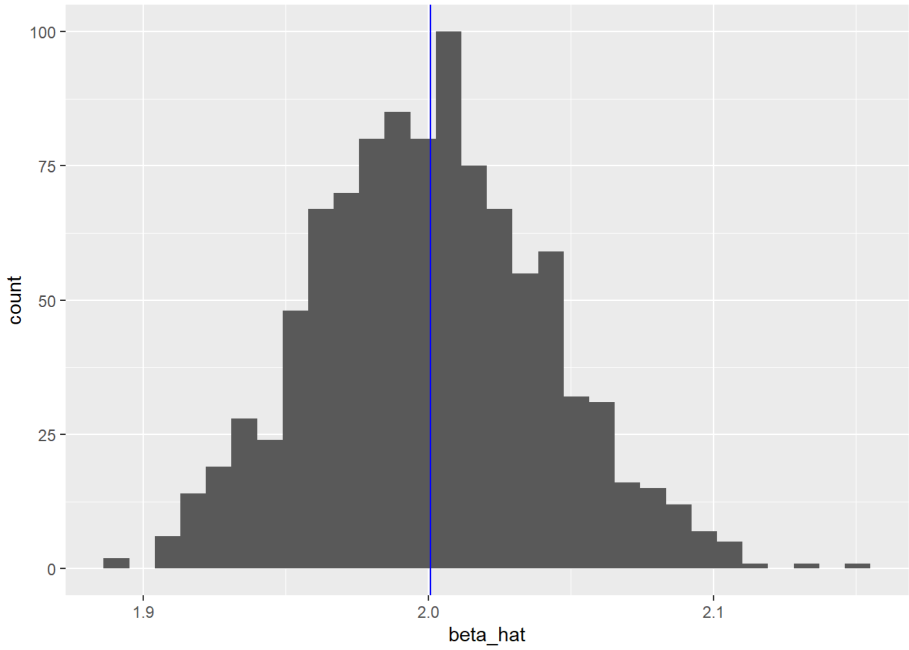

```
## the mean for beta for 1000 observations is: 2.00433538453461
## the mean for beta std for 1000 observations is: 0.122992483868974
## the real std is: 0.126701280941369
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## the mean for beta for 10000 observations is: 2.00074413739196
## the mean for beta std for 10000 observations is: 0.0416474627423592
## the real std is: 0.0397672608236585
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



we can see that the average, std and histograms change with different n values, as n increases

we can see that the average, std and histograms change with different n values, as n increases $mean(\hat{\beta}_1)$ get closer to 2
also the real sigma decreaces, and the estimated sigma as well, the histograms look more symetric around the real mean and look rather
the fact that $\hat{\beta}_1$ is a consistent estimator is seen in the simultaion, as n increaces $\bar{\hat{\beta}}_1$ gets close to 2

Q3

שאלה 3

בשאלה זו נשתמש בנתונים אודות myocardial infection בהם השתמשתם בכיתה, הזמינים כקובץ csv במודל.

העמודות הנדרשות לנו בשאלה זו הן -

- Age - גיל בו לקה/לקתה באוטם ראשון בשריר הלב (נומרי)
- CVDeath_2012 - האם נפטר מסיבות לבביות לאחר אוטם ראשון בשריר הלב (0 = לא, 1 = כן)

בתרגיל 8, התאמתם מודל רגרסיה לוגיסטית לתמותה מסיבות לבביות לאחר אוטם ראשון בשריר הלב כפונקצייה של הגיל בו לקה/לקתה באוטם ראשון בשריר הלב. קיבלתם אומדים $\hat{\beta}_0, \hat{\beta}_1$ לחותך ולגיל שחושבו בעזרת שיטה נומרית בשם Fisher Scoring.

בשאלה הזאת תנסו לשחזר את אמידת המקדמים בעצמכם.

א. הסבירו מדוע במודל זה אנו זקוקים לשיטות נומריות על מנת למצוא אומדי נראות מקסימליים.

ב. כיתבו פונקציה המקבלת $\hat{\beta}_0, \hat{\beta}_1$ כלשהם ומחזירה את לוג הנראות של מודל זה עבור אומדים אלו.

ג. כעת נבחן 100 ערכים פוטנציאליים לכל אחד מבין $\hat{\beta}_0, \hat{\beta}_1$ ונחפש את הצירוף שממקסם את לוג הנראות.
נבחר 100 ערכים במרווחים שווים בין 5- לבין 5 עבור $\hat{\beta}_0$,
נבחר 100 ערכים במרווחים שווים בין 5- לבין 5 עבור $\hat{\beta}_1$,
בדקו מה ערך לוג הנראות בכל אחד מבין הצירופים של הערכים הללו, והחזירו את צירוף ה $\hat{\beta}_0, \hat{\beta}_1$ עבורם לוג הנראות מקבל את הערך הגדול ביותר (מבין כל הצירופים שנבדקו).

ד. כעת נבחן 500 ערכים פוטנציאליים לכל אחד מבין $\hat{\beta}_0, \hat{\beta}_1$ ונחפש את הצירוף שממקסם את לוג הנראות.
נבחר 500 ערכים במרווחים שווים בין 5- לבין 5 עבור $\hat{\beta}_0$,
נבחר 500 ערכים במרווחים שווים בין 5- לבין 5 עבור $\hat{\beta}_1$,
בדקו מה ערך לוג הנראות בכל אחד מבין הצירופים של הערכים הללו, והחזירו את צירוף ה $\hat{\beta}_0, \hat{\beta}_1$ עבורם לוג הנראות מקבל את הערך הגדול ביותר (מבין כל הצירופים שנבדקו).

ה. השוו את האומדים שהתקבלו בשני הסעיפים הקודמים אחד לשני וגם לאומד שהתקבל מהתאמת המודל.

a

there is no closed form for the MLE therefore we need to find the maximum numericlly

b

```
MI <- read.csv("MI_PracticeDataset.csv") %>%
  select(Age,CVDeath_2012)

ML_logistic <- function(beta_0,beta_1){
  p <-  expit(beta_0 + beta_1 * MI$Age)
  y <- MI$CVDeath_2012
  log_lik <- as.numeric(t(y)%*%log(p) + t(1-y)%*%log(1-p))
  return(log_lik)
}
```

c

```
beta_0_canidates <- seq(-5,5,length.out=100)
beta_1_canidates <- seq(-5,5,length.out=100)
search_grid <- expand.grid(beta_0 = beta_0_canidates,beta_1 = beta_1_canidates)
glue("best values in grid are:")
```

```
## best values in grid are:
```

```
val_100 <- search_grid[which.max(map2_dbl(search_grid$beta_0,search_grid$beta_1,ML_logistic)),]
val_100
```

```
##         beta_0     beta_1
## 5010 -4.090909 0.05050505
```

d

```
beta_0_canidates <- seq(-5,5,length.out=500)
beta_1_canidates <- seq(-5,5,length.out=500)
search_grid <- expand.grid(beta_0 = beta_0_canidates,beta_1 = beta_1_canidates)
glue("best values in grid are:")
```

```
## best values in grid are:
```

```
val_500 <- search_grid[which.max(map2_dbl(search_grid$beta_0,search_grid$beta_1,ML_logistic)),]
val_500
```

```
##           beta_0    beta_1
## 126046 -4.098196 0.0501002
```

e

```
model_c <- glm(CVDeath_2012~Age,data = MI,family = "binomial")
coef_mat_c <- tidy(model_c)
beta_c <- coef_mat_c$estimate
out_mat <- cbind(c("simulation 100","simulation 500","glm"),rbind(val_100,val_500,beta_c))
colnames(out_mat) <- c("method","b0","b1")
rownames(out_mat) <- 1:3
out_mat
```

```
##           method        b0         b1
## 1 simulation 100 -4.090909 0.05050505
## 2 simulation 500 -4.098196 0.05010020
## 3            glm -4.063376 0.04941359
```

values are rather close

Q4

שאלה 4

בשאלה זו נשתמש בנתונים אודות myocardial infection בהם השתמשתם בכיתה, הזמינים כקובץ csv במודל.

העמודות הנדרשות לנו בשאלה זו הן -

- Sex - מגדר (1 = גבר, 2 = אישה)
- Age - גיל בו לקה/לקתה באוטם ראשון בשריר הלב (נומרי)
- cursmoker - האם מעשנ/ת כעת (0 = לא, 1 = כן)
- CVDeath_2012 - האם נפטר מסיבות לבביות לאחר אוטם ראשון בשריר הלב (0 = לא, 1 = כן)

א. התאימו מודל רגרסיה לוגיסטית לתמותה מסיבות לבביות לאחר אוטם ראשון בשריר הלב כפונקצייה של מגדר ועישון והאינטראקציה ביניהם.

I. דווחו את התוצאות שקיבלתם. מה המסקנות?
II. חשבו את ה Pearson residuals של כל תצפית לפי המודל הזה וציירו אותן בגרף מתאים. מה ניתן ללמוד מהן?
III. חשבו את ה Deviance residuals של כל תצפית לפי המודל הזה וציירו אותן בגרף מתאים. מה ניתן ללמוד מהן?
IV. חשבו את האומד לOR המתקבל ממודל זה, בין גבר שלא מעשן לבין אישה שמעשנת.
V. חשבו רווח סמך ברמת סמך 95% לOR המתקבל ממודל, זה בין גבר שלא מעשן לבין אישה שמעשנת.

ב. התאימו מודל רגרסיה לוגיסטית לתמותה מסיבות לבביות לאחר אוטם ראשון בשריר הלב כפונקצייה של גיל ועישון והאינטראקציה ביניהם.

I. דווחו את התוצאות שקיבלתם. מה המסקנות?
II. חשבו את ה Pearson residuals של כל תצפית לפי המודל הזה וציירו אותן בגרף מתאים. מה ניתן ללמוד מהן?
III. חשבו את ה Deviance residuals של כל תצפית לפי המודל הזה וציירו אותן בגרף מתאים. מה ניתן ללמוד מהן?
IV. דווחו את האומד לOR המתקבל ממודל זה, בין אדם בן 50 שלא מעשן לבין אדם שמעשן.
V. חשבו רווח סמך ברמת סמך 95% לOR המתקבל ממודל זה, בין אדם בן 50 שלא מעשן לבין אדם שמעשן.
VI. דווחו את האומד לOR המתקבל ממודל זה, בין אדם בן 50 שלא מעשן לבין אדם שמעשן ומבוגר ממנו ב7 שנים.
VII. חשבו רווח סמך ברמת סמך 95% לOR המתקבל ממודל זה, בין אדם בן 50 שלא מעשן לבין אדם שמעשן ומבוגר ממנו ב7 שנים.

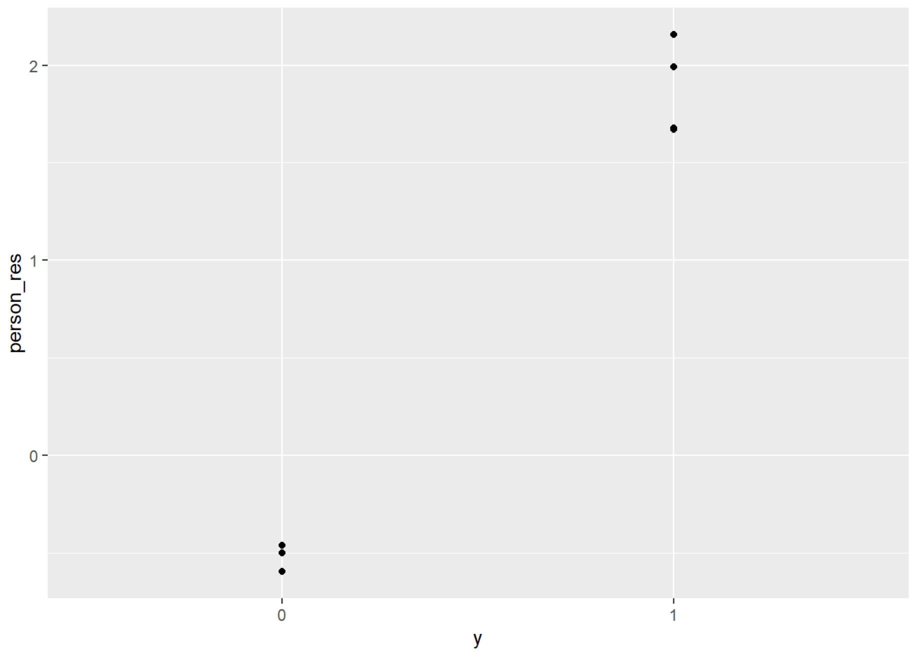
a

```
MI <- read.csv("MI_PracticeDataset.csv") %>%
  select(cursmoker,Sex,Age,CVDeath_2012) %>%
  mutate(across(c(cursmoker,Sex,CVDeath_2012),factor))
y = MI$CVDeath_2012
model <- glm(CVDeath_2012~cursmoker + Sex+Sex*cursmoker,data = MI,family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = CVDeath_2012 ~ cursmoker + Sex + Sex * cursmoker,
##     family = "binomial", data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7820  -0.6707  -0.6707  -0.6240   1.8613
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.5374     0.1131 -13.595   <2e-16 ***
## cursmoker1         0.1601     0.1474   1.086   0.2775
## Sex2               0.5093     0.2012   2.531   0.0114 *
## cursmoker1:Sex2   -0.1668     0.3157  -0.529   0.5971
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1540.9  on 1520  degrees of freedom
## Residual deviance: 1532.5  on 1517  degrees of freedom
## AIC: 1540.5
##
## Number of Fisher Scoring iterations: 4
```

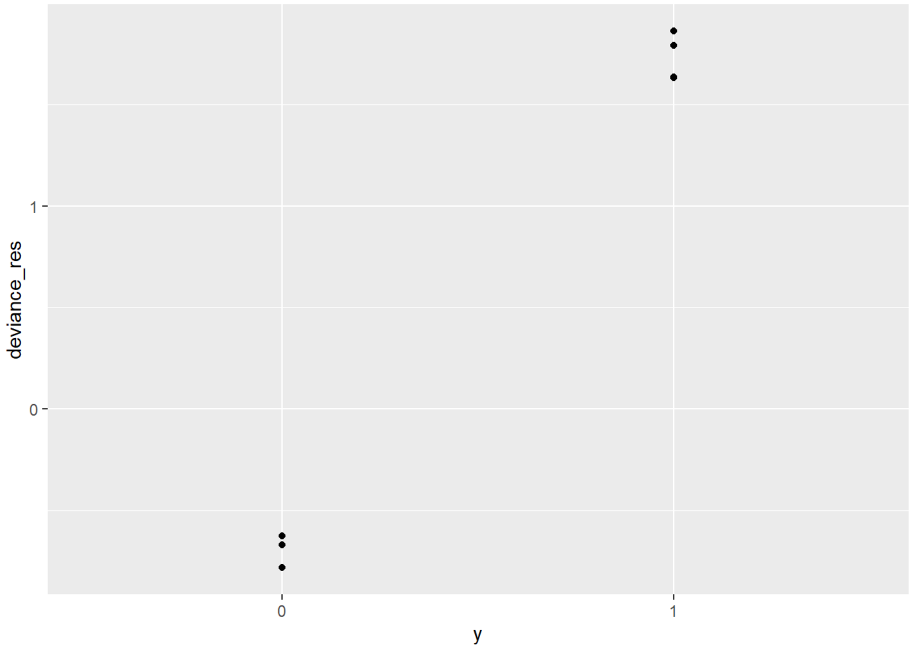
sex has a large impact on the model but cursmoker and the interaction is not significant

```
person_res <- residuals(model, type = "pearson")

tibble(y = y,person_res = person_res)%>%
  ggplot(aes(x = y,y = person_res))+
    geom_point()
```



we can learn that given y=1 the variance of pearson residuals is larger then y =0

```
deviance_res <- residuals(model, type = "deviance")

tibble(y = y,deviance_res = deviance_res)%>%
  ggplot(aes(x = y,y = deviance_res))+
    geom_point()
```



we can learn that given y=1 the variance of deviance residuals is the same as y =0

```
coef_mat <- tidy(model)
beta <- coef_mat$estimate
diff_vec <- c(1,0,0,0) - c(1,1,1,1)
pe <- beta%*% diff_vec
OR <- as.numeric(exp(pe))
v_or <- as.numeric(sqrt(diff_vec %*% vcov(model) %*% diff_vec))

glue("the OR estimate is: {round(OR,3)}
     and the CI is ({round(exp(pe -  qnorm(0.975)*v_or),3)},{round(exp(pe +qnorm(0.975)*v_or),3)})")
```

```
## the OR estimate is: 0.605
## and the CI is (0.37,0.989)
```

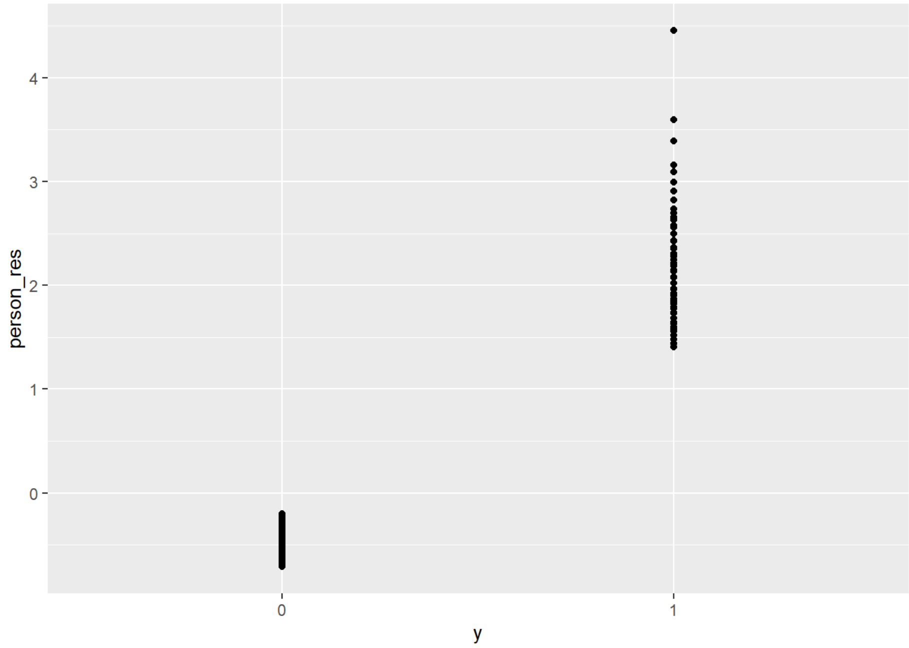
b

```
model_b <- glm(CVDeath_2012~cursmoker + Age+cursmoker*Age,data = MI,family = "binomial")
summary(model_b)
```

```
##
## Call:
## glm(formula = CVDeath_2012 ~ cursmoker + Age + cursmoker * Age,
##     family = "binomial", data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9062  -0.7380  -0.6159  -0.4599   2.4643
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -4.873798   0.886385  -5.499 3.83e-08 ***
## cursmoker1      0.798798   1.077193   0.742    0.458
## Age             0.060857   0.015172   4.011 6.04e-05 ***
## cursmoker1:Age -0.008591   0.018927  -0.454    0.650
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1540.9  on 1520  degrees of freedom
## Residual deviance: 1499.5  on 1517  degrees of freedom
## AIC: 1507.5
##
## Number of Fisher Scoring iterations: 4
```

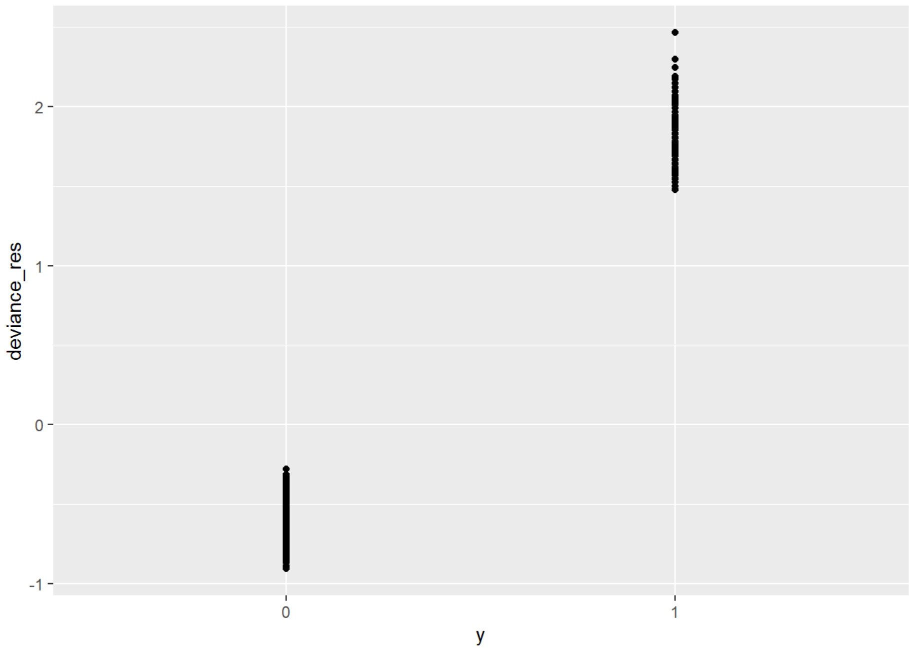
age has a large impact on the model but cursmoker and the interaction is not significant

```
person_res <- residuals(model_b, type = "pearson")

tibble(y = y,person_res = person_res)%>%
  ggplot(aes(x = y,y = person_res))+
    geom_point()
```



we can learn that given y=1 the variance of pearson residuals is larger then y =0

```
deviance_res <- residuals(model_b, type = "deviance")

tibble(y = y,deviance_res = deviance_res)%>%
  ggplot(aes(x = y,y = deviance_res))+
    geom_point()
```



we can learn that given y=1 the variance of deviance residuals is the same as y =0

```
coef_mat <- tidy(model_b)
beta <- coef_mat$estimate
diff_vec <- c(1,0,50,0) - c(1,1,50,50)
pe <- beta%*% diff_vec
OR <- as.numeric(exp(pe))
v_or <- as.numeric(sqrt(diff_vec %*% vcov(model_b) %*% diff_vec))

glue("the OR for a 50 year old person who doesnt smoke or smoke estimate is: {round(OR,3)}
     and the CI is ({round(exp(pe -  qnorm(0.975)*v_or),3)},{round(exp(pe +qnorm(0.975)*v_or),3)})")
```

```
## the OR for a 50 year old person who doesnt smoke or smoke estimate is: 0.691
## and the CI is (0.483,0.989)
```

```
coef_mat <- tidy(model_b)
beta <- coef_mat$estimate
diff_vec <- c(1,0,50,0) - c(1,1,57,57)
pe <- beta%*% diff_vec
OR <- as.numeric(exp(pe))
v_or <- as.numeric(sqrt(diff_vec %*% vcov(model_b) %*% diff_vec))

glue("the OR for a 50 year old person who doesnt smoke and a 57 year old person who smokes estimate is: {round(OR,3)}
     and the CI is ({round(exp(pe -  qnorm(0.975)*v_or),3)},{round(exp(pe +qnorm(0.975)*v_or),3)})")
```

```
## the OR for a 50 year old person who doesnt smoke and a 57 year old person who smokes estimate is: 0.479
## and the CI is (0.335,0.686)
```