EX8

roi hezkiyahu

28 4 2022

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

Q1

שאלה 1

בשאלה זו נבחן את פונקציית ה $expit$ ואת פונקציית ה $logit$.

תזכורת:

- $expit(u) = \frac{e^u}{1+e^u}$
- $logit(p) = log\frac{p}{1-p}$

א. הוכיחו כי פונקציית ה $expit$ היא פונקציה מונוטונית עולה.

ב. הוכיחו כי פונקציית ה $logit$ היא פונקציה מונוטונית עולה.

ג. הוכיחו כי כאשר p גדל, ההפרש בין פונקציית ה $logit$ לבין פונקציית ה log גדל.

ד. בסעיף זה נצייר את ערכי פונקציית ה $logit$ וגם את ערכי פונקציית ה log עבור אותם ערכי p, ונשווה בין הציורים.
- צרו 1000 ערכי p שונים בין 0 ל1 (לא כולל)
- לכל ערך p, חשבו את ערך ה $logit$ המתאים וגם את ערך ה log המתאים
- ציירו גרף של ה $logit$ וגם של ה log כפונקצייה של p.
מה ניתן ללמוד מהגרף? האם זה תואם את המסקנה שלכם מסעיף ג'?

ה. כתבו פונקציה המקבלת ערך $d \geq 0$ ומחזירה את ערך ה p הגדול ביותר אשר מבטיח שההפרש בין פונקציית ה $logit$ לבין פונקציית ה log יהיה קטן או שווה לערך d.
- נדרש דיוק של 3 ספרות אחרי הנקודה.

a

$$\text{assume } v > u$$
$$expit(v) = \frac{e^v}{1+e^v} > \frac{e^u}{1+e^u} = expit(u) \iff e^v + e^{v+u} > e^u + e^{v+u} \iff e^v > e^u \iff v > u$$

b

$$p1 > p2 logit(p1) = log(\frac{p1}{1-p1}) > log(\frac{p2}{1-p2}) = logit(p2) \iff log(p1) - log(1-p1) > log(p2) - log(1-p2) \iff log(p1) - log(p2) >$$
$$\iff log(\frac{p1}{p2}) > log(\frac{1-p1}{1-p2}) \iff \frac{p1}{p2} > \frac{1-p1}{1-p2}$$
$$\text{the last equality holds because: } \frac{p1}{p2} > 1, \frac{1-p1}{1-p2} < 1$$

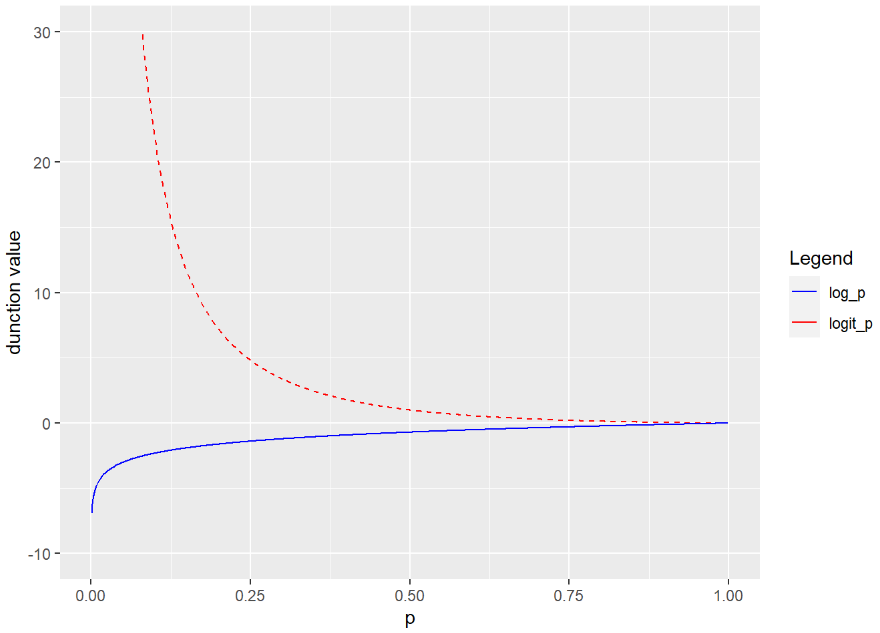
c

$$\text{we need to show that } |log(p) - logit(p)| \text{ is a monotonic increasing function}$$
$$|log(p) - logit(p)| = |log(p) - log(p) + log(1-p)| = |log(1-p)|$$
$$\text{which is a monotonic increasing function of p}$$

d

```
#logit function
logit <- function(p){log(p)/log(1-p)}
#create values
p_values <- seq(0,1,length.out=1002)[2:1001]
#calculate log values
log_p <- log(p_values)
#calculate logit values
logit_p <- logit(p_values)
tbl <- tibble("p"=p_values,
              "log_p"= log_p,
              "logit_p" = logit_p)
tbl %>%
  ggplot(aes(x = p_values,y=logit_p,z = log_p))+
  geom_line(aes(x=p_values,y=logit_p,color = "logit_p"),lty = 2)+
  geom_line(aes(x=p_values,y=log_p,color = "log_p"))+
  labs(x = "p",
       y = "dunction value",
       color = "Legend")+
  scale_color_manual(values = c("logit_p" = "red","log_p" = "blue"))+
  ylim(-10,30)
```

```
## Warning: Removed 80 row(s) containing missing values (geom_path).
```



we can see from the graph that when p increases the distance between the 2 functions decreases and the lines are rather close from p>0.5, this is exactly our conclusion on the last question

e

```
d_dist <- function(d){
  potienal_p <- seq(0.0000005,1,0.0000005)
  for (i in (length(potienal_p)-1):1) {
    p = potienal_p[i]
    if (logit(p)-log(p)>d){
      return(potienal_p[i+1])
    }
  }
  return(potienal_p[1])
}
```

Q2

שאלה 2

נתון מודל רגרסיה לוגיסטית,

$$\Pr(Y=1|X=x) = expit(\beta^T x)$$

עם וקטור משתנים מסבירים $X=(1,x_1,...,x_K)$.

הוכיחו כי e^{β_k} שווה ל OR בין שתי תצפיות/קבוצות בעלות אותם ערכי X, מלבד x_k, וכן ש x_k בתצפית/קבוצה אחת גדול ביחידה אחת לעומת התצפית/קבוצה השניה.

$$x = (1, x_1, \ldots, x_k)$$
$$y = (1, x_1, \ldots, x_k + 1)$$
$$expit(\beta^t x) = \frac{e^{\beta^t x}}{1+e^{\beta^t x}} = \frac{e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k}}{1+e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k}}$$
$$expit(\beta^t y) = \frac{e^{\beta^t y}}{1+e^{\beta^t y}} = \frac{e^{\beta^t_{-k} y_{-k}} e^{\beta_k y_k}}{1+e^{\beta^t_{-k} y_{-k}} e^{\beta_k y_k}} = \frac{e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k} e^{\beta_k}}{1+e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k} e^{\beta_k}}$$
$$\frac{Pr(Y=1|X=x)}{1-Pr(Y=1|X=x)} = \frac{expit(\beta^t x)}{1-expit(\beta^t x)} = \frac{e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k}}{1+e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k}} \Big/ \frac{1}{1+e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k}} = e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k}$$
$$\frac{Pr(Y=1|X=y)}{1-Pr(Y=1|X=y)} = \ldots \text{same as above} \ldots = e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k} e^{\beta_k}$$
$$OR = \frac{\frac{Pr(Y=1|X=x)}{1-Pr(Y=1|X=x)}}{\frac{Pr(Y=1|X=y)}{1-Pr(Y=1|X=y)}} = \frac{e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k}}{e^{\beta^t_{-k} x_{-k}} e^{\beta_k x_k} e^{\beta_k}} = e^{\beta_k}$$

Q3

שאלה 3

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)
- CVDeath_2012 - האם נפטר מסיבות לבביות לאחר אוטם ראשון בשריר הלב (0 = לא, 1 = כן)

נרצה לבחון את הקשר בין מגדר לבין תמותה מסיבות לבביות לאחר אוטם ראשון בשריר הלב. לשם כך,

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

I. האם זהו מודל רווי? הסבירו.
II. דווחו את התוצאות שקיבלתם. מה המסקנות?
III. דווחו את האומד לOR שהתקבל מהמודל.
הסבירו במילים מה המשמעות של האומד שהתקבל.
האם הוא זהה ל OR שחישבתם בתרגיל 7 שאלה 3?
IV. חשבו רווח סמך ברמת סמך של 95% לOR בעזרת ההתפלגות האסימפטוטית של האומדים, וגם בעזרת הפונקציה confint.default(). השוו בין רווחי הסמך שהתקבלו, האם הם זהים?

ב. התאימו מודל רגרסיה לוגיסטית לתמותה מסיבות לבביות לאחר אוטם ראשון בשריר הלב כפונקצייה של הגיל בו לקה/לקתה באוטם ראשון בשריר הלב.

I. האם זהו מודל רווי? הסבירו.
II. דווחו את התוצאות שקיבלתם. מה המסקנות?
III. דווחו את האומד לOR שהתקבל מהמודל.
הסבירו במילים מה המשמעות של האומד שהתקבל.
IV. חשבו רווח סמך לOR ברמת סמך של 95% בעזרת ההתפלגות האסימפטוטית של האומדים, וגם בעזרת הפונקציה confint.default(). השוו בין רווחי הסמך שהתקבלו, האם הם זהים?
V. חשבו אומד נקודתי ורווח סמך ברמת סמך של 95% לOR המשווה בין גיל 40 לגיל 50.
VI. חשבו אומד נקודתי ורווח סמך ברמת סמך של 95% לOR המשווה בין גיל 50 לגיל 60.
VII. עבור אלו שלקו באוטם ראשון בשריר הלב בגיל 50,
(1) חשבו אומד נקודתי לסיכוי לתמותה בעקבות אוטם ראשון בשריר הלב.
(2) כתבו את הערכים של $x^T, \hat{\beta}, \hat{\Sigma}$ המתקבלים ממודל זה במדגם זה.
(3) בשיעור ראינו כי רווח סמך ברמת סמך של 95% לאומד לסיכוי מתקבל על ידי

$$expit(x^T\hat{\beta} \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{x^T\hat{\Sigma}x})$$

השתמשו בנוסחה זו ובאומדים שחישבתם בסעיף הקודם וחשבו רווח סמך ברמת סמך של 95% לאומד לסיכוי לתמותה בעקבות אוטם ראשון בשריר הלב.

ג. התאימו מודל רגרסיה לוגיסטית לתמותה מסיבות לבביות לאחר אוטם ראשון בשריר הלב כפונקצייה של גיל, ושל גיל בריבוע.

I. דווחו את התוצאות שקיבלתם. מה המסקנות?
II. חשבו אומד נקודתי ורווח סמך לOR המשווה בין גיל 40 לגיל 50.
III. חשבו אומד נקודתי ורווח סמך לOR המשווה בין גיל 50 לגיל 60.

a

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

```
##
## Call:
## glm(formula = CVDeath_2012 ~ Sex, family = "binomial", data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7812  -0.6506  -0.6506  -0.6506   1.8203
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.44517    0.07251 -19.929  < 2e-16 ***
## Sex2         0.41461    0.15202   2.727  0.00639 **
## ---
## 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: 1533.7  on 1519  degrees of freedom
## AIC: 1537.7
##
## Number of Fisher Scoring iterations: 4
```

the model is not saturated because we have more data points then parameters

we can see that sex has a large effect on CVDeath_2012, where woman has a higher chance to die

$$OR = e^{\beta_1} = e^{0.41461} \approx 1.51378, \text{it is the same as the previous exercise results}$$

the OR estimates suggests that a man chance of dying from myocardial infection compared to not getting an infection is 151% of the chance woman has

```
conf_def <- exp(confint.default(model)[2,])
names(conf_def) <- c()
#tidy coef matrix
coef_mat <- tidy(model)
conf_glm <- exp(coef_mat$estimate[2] + c(-1,1) * qnorm(0.975)* coef_mat$std.error[2])
tibble("method"= c("confint.default","glm"),"L"= c(conf_def[1],conf_glm[1]),"U"= c(conf_def[2],conf_glm[2]))
```

```
## # A tibble: 2 x 3
##   method              L     U
##   <chr>           <dbl> <dbl>
## 1 confint.default  1.12  2.04
## 2 glm              1.12  2.04
```

```
#is it close?
near(conf_def,conf_glm)
```

```
## [1] TRUE TRUE
```

the CIs are the same

b

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

```
##
## Call:
## glm(formula = CVDeath_2012 ~ Age, family = "binomial", data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8431  -0.7265  -0.6223  -0.4725   2.3238
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.063376   0.484489  -8.387  < 2e-16 ***
## Age          0.049414   0.008627   5.728 1.02e-08 ***
## ---
## 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: 1505.2  on 1519  degrees of freedom
## AIC: 1509.2
##
## Number of Fisher Scoring iterations: 4
```

the model is not saturated because we have more data points then parameters *we can see that higher age groups have a higher chance of not surviving CV*

$$OR = e^{\beta_1} = e^{0.049414} \approx 1.05$$

the OR estimates suggests that getting older by one year rasies the chance of dying from myocardial infection compared to not getting an infection by 5%

```
conf_def2 <- exp(confint.default(model_b)[2,])
names(conf_def2) <- c()
#tidy coef matrix
coef_mat2 <- tidy(model_b)
conf_glm2 <- exp(coef_mat2$estimate[2] + c(-1,1) * qnorm(0.975)* coef_mat2$std.error[2])
tibble("method"= c("confint.default","glm"),"L"= c(conf_def2[1],conf_glm2[1]),"U"= c(conf_def2[2],conf_glm2[2]))
```

```
## # A tibble: 2 x 3
##   method              L     U
##   <chr>           <dbl> <dbl>
## 1 confint.default  1.03  1.07
## 2 glm              1.03  1.07
```

```
#is it close?
near(conf_def2,conf_glm2)
```

```
## [1] TRUE TRUE
```

$$\text{the estimate for the OR between 40 and 50 years is the same as the OR between 50 and 60 and is: } e^{10\beta 1}$$
$$\beta_1 \sim N(0.049414, 0.008627) \Rightarrow 10\beta_1 \sim N(0.49414, 0.8627)$$

```
glue("the OR estimate is: {round(exp(0.49),3)}
     and the CI is ({round(exp(0.49 -  qnorm(0.975)*0.8627),3)},{round(exp(0.49 +qnorm(0.975)*0.8627),3)})")
```

```
## the OR estimate is: 1.632
## and the CI is (0.301,8.854)
```

```
expit <- function(p){exp(p)/(1+exp(p))}
x <- c(1,50)
beta_hat <-  coef_mat2$estimate
x_t_beta <- t(x)%*%beta_hat
glue("for a 50 year old the chance of dying is: {round(expit(x_t_beta)*100,2)}%")
```

```
## for a 50 year old the chance of dying is: 16.9%
```

```
sigma_hat <- vcov(model_b)

glue("sigma hat is:")
```

```
## sigma hat is:
```

```
sigma_hat
```

```
##              (Intercept)           Age
## (Intercept)  0.234729135 -4.142683e-03
## Age         -0.004142683  7.442445e-05
```

```
glue("beta hat is:{beta_hat[1]},{beta_hat[2]}")
```

```
## beta hat is:-4.06337605790547,0.0494135907652687
```

```
c_i <-  round(expit(as.numeric(x_t_beta) + c(-1,1)* qnorm(0.975) * as.numeric(sqrt(t(x)%*%sigma_hat%*%x))),4)
glue("CI for dying chance is: ({c_i[1]*100}%,{c_i[2]*100}%)")
```

```
## CI for dying chance is: (14.79%,19.24%)
```

c

```
MI <- MI %>%
  mutate(Age_saqured = Age^2)
model_c<- glm(CVDeath_2012~Age + Age_saqured ,data = MI,family = "binomial")
summary(model_c)
```

```
##
## Call:
## glm(formula = CVDeath_2012 ~ Age + Age_saqured, family = "binomial",
##     data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8743  -0.7150  -0.6047  -0.4987   2.1269
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.4192270  2.5401619  -0.559    0.576
## Age         -0.0537227  0.0981165  -0.548    0.584
## Age_saqured  0.0009814  0.0009330   1.052    0.293
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1540.9  on 1520  degrees of freedom
## Residual deviance: 1504.1  on 1518  degrees of freedom
## AIC: 1510.1
##
## Number of Fisher Scoring iterations: 4
```

we can see that: age^2 is non significant

```
coef_mat_c <- tidy(model_c)
beta_hat_c <- coef_mat_c$estimate
diff_vec_40_50 <- c(10,50^2-40^2)
diff_vec_50_60 <- c(10,60^2-50^2)
#point estimates
pe_40_50 <- as.numeric(beta_hat_c[c(2,3)]%*%diff_vec_40_50)
#point estimates
pe_50_60 <- as.numeric(beta_hat_c[c(2,3)]%*%diff_vec_50_60)
#variance estimation:
v_40_50 <- as.numeric(sqrt(diff_vec_40_50 %*% vcov(model_c)[c(2,3),c(2,3)] %*% diff_vec_40_50))
v_50_60 <- as.numeric(sqrt(diff_vec_50_60 %*% vcov(model_c)[c(2,3),c(2,3)] %*% diff_vec_50_60))
glue("the OR estimate for comparing age 40 to 50 is: {round(exp(pe_40_50),3)}
     and the CI is ({round(exp(pe_40_50 -  qnorm(0.975)*v_40_50),3)},{round(exp(pe_40_50 +qnorm(0.975)*v_40_50),3)})
     the OR estimate for comparing age 50 to 60 is: {round(exp(pe_50_60),3)}
     and the CI is ({round(exp(pe_50_60 -  qnorm(0.975)*v_50_60),3)},{round(exp(pe_50_60 +qnorm(0.975)*v_50_60),3)})")
```

```
## the OR estimate for comparing age 40 to 50 is: 1.413
## and the CI is (1.03,1.939)
## the OR estimate for comparing age 50 to 60 is: 1.72
## and the CI is (1.422,2.08)
```