EX7

roi hezkiyahu

19 4 2022

```
# imports
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
------------------------------------------------------------------------------- tidyverse 1.3.0 --
```

```
## <U+221A> ggplot2 3.3.2     <U+221A> purrr   0.3.4
## <U+221A> tibble  3.0.3     <U+221A> dplyr   1.0.2
## <U+221A> tidyr   1.1.2     <U+221A> stringr 1.4.0
## <U+221A> readr   1.3.1     <U+221A> forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
------------------------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(glue)
```

```
##
## Attaching package: 'glue'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(lmerTest)
```

```
## Loading required package: lme4
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
##
## Attaching package: 'lmerTest'
```

```
## The following object is masked from 'package:lme4':
##
##     lmer
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
library(lme4)
```

Q1

שאלה 1

בשאלה זו נשתמש בנתונים בשם feedback_bi הזמינים כקובץ csv במודל.

בתרגיל בית 5 שאלה 4 סעיף ד', התאמתם מודל לנתונים אלה, עם חותך מקרי למשתתפים השונים (id), ועם מגדר (gender), פידבק (feedback) והאינטראקציה בניהם כאפקטים קבועים. המשתנה התלוי היה התיפקוד (performance).

א. נרצה לבדוק בעזרת GLRT אילו משתנים מובהקים,

I. בידקו בעזרת GLRT האם האפקט המקרי מובהק עבור $\alpha = 0.05$.

II. בידקו בעזרת GLRT האם האינטראקציה של המגדר (gender) והפידבק (feedback) מובהקת עבור $\alpha = 0.05$.

III. כיצד מתפלג אסימפטוטית $log\Lambda$ תחת השערת האפס בכל אחד מהסעיפים? הסבירו.

ב. במצגת 7 שקף 14 מוצגת הדרך הבאה לאמוד את b_i:

$$\widehat{b_i} = \widehat{B} Z_i^T \widehat{\Sigma_i}^{-1} (Y_i - X_i \widehat{\beta})$$

כעת בכל סעיף התמקדו בקלאסטר ספציפי, אמדו את b_i בעזרת הנוסחה והשוו לאומד של b_i שמתקבל מהמודל (תוכלו להיעזר בפונקציה coef). האם קיבלתם אותו הדבר בשתי השיטות (הסתפקו בדיוק של ספרה אחת)?

I. חשבו ידנית את האומד ל b_i והשוו אותו לאומד מהמודל, עבור id=7.

II. חשבו ידנית את האומד ל b_i והשוו אותו לאומד מהמודל, עבור id=12.

III. מה ניתן לומר על התפקוד של id=7 לעומת התפקוד של id=12 ?

a

```
feed_back <- read.csv("feedback_df_bi.csv") %>%
  select(id,performance,feedback,gender) %>%
  mutate(across(c(gender,feedback),as.factor))
#full model
model <- lmer(performance ~ feedback+ gender+ gender*feedback + (1|id) , data = feed_back)
#no random effect model
model_no_ra <- lm(performance ~ feedback+ gender+ gender*feedback , data = feed_back)
anova(model,model_no_ra)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: feed_back
## Models:
## model_no_ra: performance ~ feedback + gender + gender * feedback
## model: performance ~ feedback + gender + gender * feedback + (1 | id)
##             npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model_no_ra    7 6441.5 6474.3 -3213.8   6427.5
## model          8 6118.9 6156.3 -3051.4   6102.9 324.65  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glue("alpha < 0.05 therefore we can reject the GLRT null and conclude that the random effect is signi
ficant")
```

```
## alpha < 0.05 therefore we can reject the GLRT null and conclude that the random effect is signific
ant
```

```
# no interaction model
model_no_inter <- lmer(performance ~ feedback+ gender+ (1|id) , data = feed_back)
anova(model,model_no_inter)
```

```
## refitting model(s) with ML (instead of REML)
```

```
## Data: feed_back
## Models:
## model_no_inter: performance ~ feedback + gender + (1 | id)
## model: performance ~ feedback + gender + gender * feedback + (1 | id)
##                npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model_no_inter    6 6137.5 6165.5 -3062.7   6125.5
## model             8 6118.9 6156.3 -3051.4   6102.9 22.573  2  1.254e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
glue("alpha < 0.05 therefore we can reject the GLRT null and conclude that the interaction is signifi
cant")
```

```
## alpha < 0.05 therefore we can reject the GLRT null and conclude that the interaction is significan
t
```

$$-2log\Lambda \xrightarrow{D} \chi^2_k \text{ where k = difference in the degrees of freedom}$$

for the random effect test the difference in the degrees of freedom is the number of participantes (22) thus:

$$log\Lambda \xrightarrow{D} \frac{-\chi^2_{22}}{2}$$

for the interaction test the difference in the degrees of freedom is 1 thus:

$$log\Lambda \xrightarrow{D} \frac{-\chi^2_{1}}{2}$$

b

```
model_summary = summary(model)
#model intercept
inter <- 98.849 #from summary
#B
var_id <- 84.91 #from summary
B <-  as.matrix(var_id)
#relevant obervations
obs7 <- feed_back$id == 7
obs12 <- feed_back$id == 12
#Z
z7 <- rep(1,36)
z12 <- rep(1,36)
#res
res7 <- feed_back$performance[obs7] - model.matrix(model)[obs7,]%*%model_summary$coefficients[,1]
res12 <- feed_back$performance[obs12] - model.matrix(model)[obs12,]%*%model_summary$coefficients[,1]
var_res <- 119.63 #from summary
#sigma
sigma7 <- z7%*%B%*%t(z7) + var_res*diag(36)
sigma12 <- z12%*%B%*%t(z12) + var_res*diag(36)

b7 <- B*t(z7)%*%solve(sigma7)%*%(res7) + inter
b12 <- B*t(z12)%*%solve(sigma12)%*%(res12) + inter

#check if distance between estimates is less than 0.01
coef(model)$id[c(7,12),1] -c(b7,b12) <0.01
```

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

we can see that the performance of subject 7 is better (by 15 units) than subject 12

Q2

שאלה 2

בשאלה זו בנתונים בשם sleepstudy מחבילת R בשם lme4

בתרגיל בית הקודם רצינו לבחון את האפקט של מספר הימים שעברו (Days) על זמן התגובה במשימה (Raction), תוך כדי שלקחנו בחשבון את השונות בין המשתתפים השונים (Subject).

לשם כך, התאמתם שני מודלים. מודל LMM עם חותך מקרי וגם מודל LMM עם חותך מקרי ושיפוע מקרי.

א. עבור כל אחד מהמודלים שהתאמתם, ציירו גרף של השאריות כפונקציה של הערכים החזויים. השוו בין שני הגרפים.

ב. עבור כל אחד מהמודלים שהתאמתם, ציירו גרף של השאריות כפונקציה של מספר הימים שעברו. השוו בין שני הגרפים.

ג. במודל LMM עם חותך מקרי,

I. מי האדם עם החותך הכי גבוה?

II. כמה אנשים יש עם חותך שלילי?

III. מה אחוז האנשים באוכלוסייה עם חותך שלילי?

ד. במודל LMM עם חותך מקרי ושיפוע מקרי,

I. מי האדם עם החותך הכי גבוה?

II. כמה אנשים יש עם חותך שלילי?

III. מה אחוז האנשים באוכלוסייה עם חותך שלילי?

IV. מי האדם עם השיפוע הכי גדול?

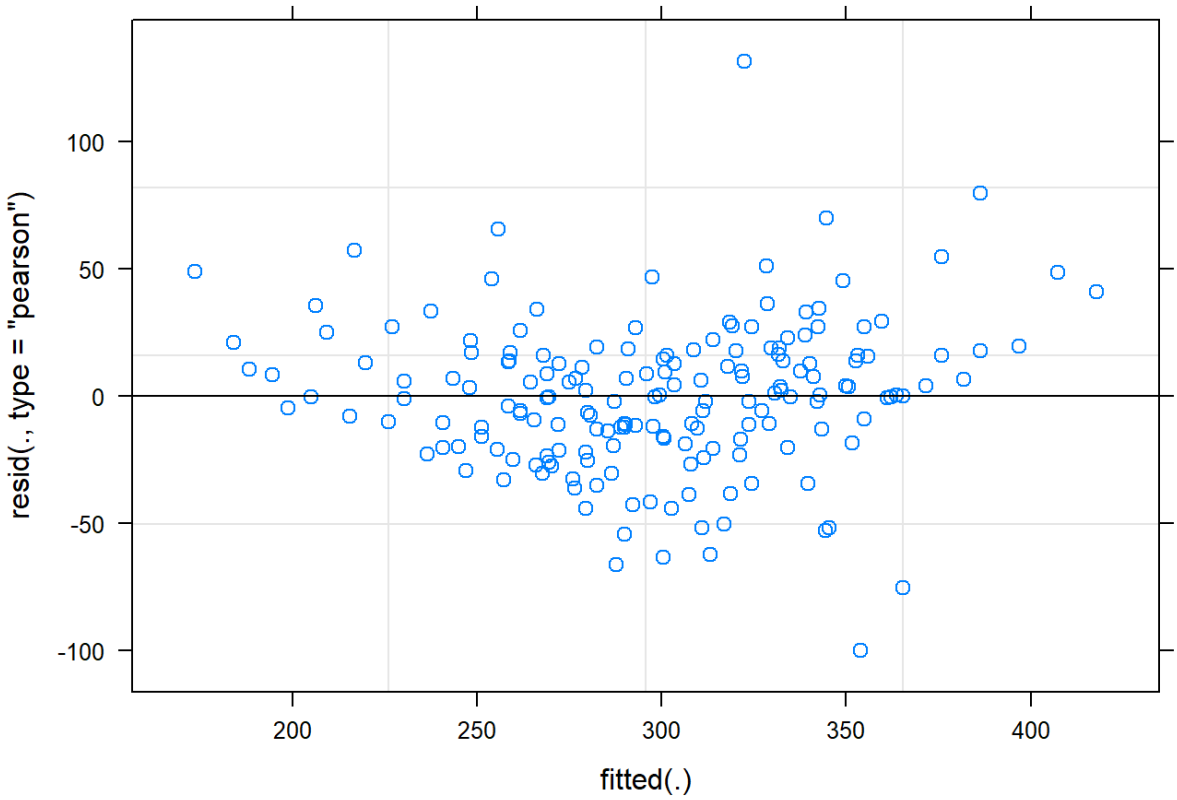
V. עבור כמה אנשים זמן התגובה נהיה מהיר יותר ככל שהימים עוברים?

VI. כמה אנשים יש עם חותך שלילי וגם עם שיפוע חיובי?

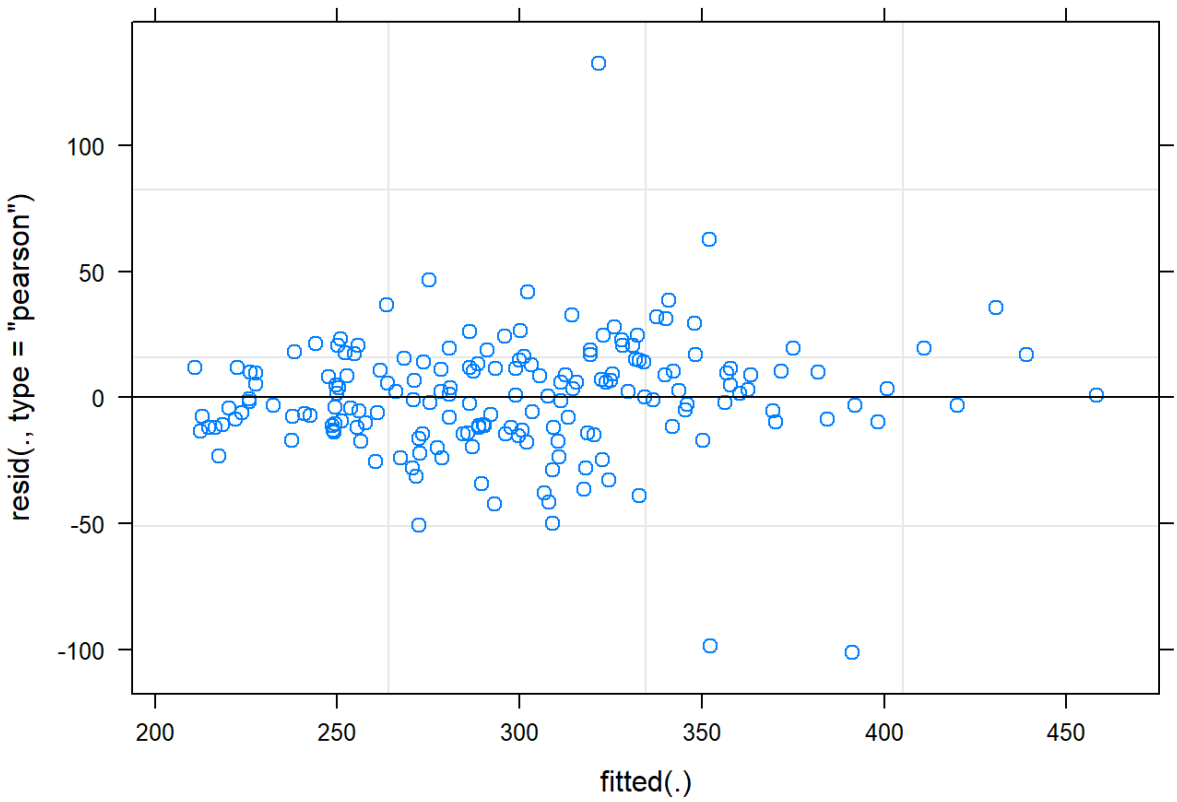
VII. כמה אנשים יש עם חותך חיובי וגם עם שיפוע שלילי?

a

```
data("sleepstudy")
model_1 <- lmer(Reaction ~ (1|Subject) + Days,data = sleepstudy)
model_2 <- lmer(Reaction ~ (1+Days|Subject) + Days,data = sleepstudy)
plot(model_1)
```



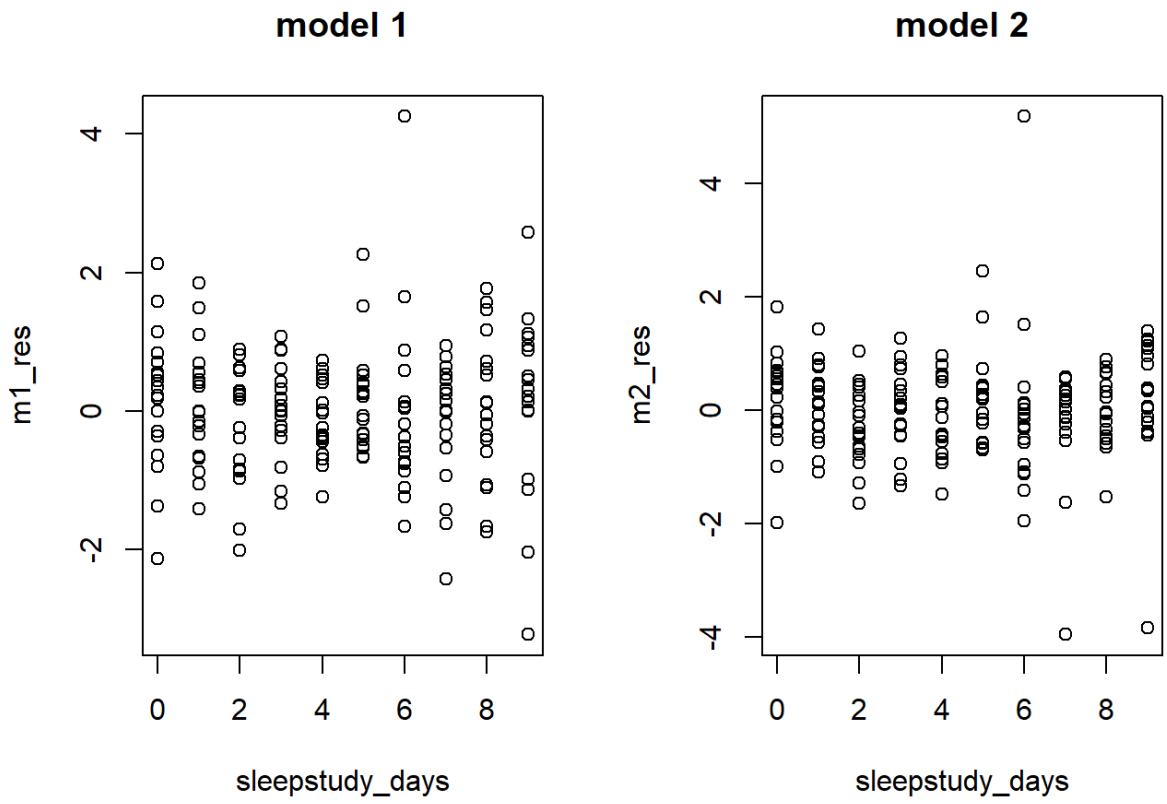

```
plot(model_2)
```



it looks like the 2nd graph better fits the data, the dots are spreaded quite even across the line, on the other hand in the first model we can see that the the residuals has a cubic form suggesting the model does not fit the data well and might need a transformation

b

```
sleepstudy_days <- sleepstudy$Days
m1_res <- summary(model_1)$residuals
m2_res <- summary(model_2)$residuals
par(mfrow = c(1,2))
plot(sleepstudy_days,m1_res, main = "model 1")
plot(sleepstudy_days,m2_res, main = "model 2")
```



model 2 has less variability than model 1

c

```
b_estimates1 <- coef(model_1)$Subject[,1] -251.41 #intercept from model summary
largest_subj1 <- which.max(b_estimates1)
neg_inter1 <- sum(b_estimates1<0)
neg_precent1 <- neg_inter1/length(b_estimates1)
glue("subject {largest_subj1} has the largest intercept, there are {neg_inter1} ({round(neg_precent1,
2)*100}%) subject with negetive intercept for model 1")
```

```
## subject 10 has the largest intercept, there are 8 (44%) subject with negetive intercept for model
1
```

d

```
b_estimates2 <- coef(model_2)$Subject[,1] -251.41#intercept from model summary
n <-  length(b_estimates2)
largest_subj2 <- which.max(b_estimates2)
lower_inter <- b_estimates2<0
neg_inter2 <- sum(lower_inter)
neg_precent2 <- neg_inter2/n
slope_estimates <- coef(model_2)$Subject[,2]
largest_slope <- which.max(slope_estimates)
faster_react <- slope_estimates > 0
neg_inetr_pos_slope <- sum(lower_inter * faster_react)
pos_inetr_neg_slope <- sum((1-lower_inter) * (1-faster_react))
glue("subject {largest_subj2} has the largest intercept,
     there are {neg_inter2} ({round(neg_precent2,2)*100}%) subject with negative intercept,
     subject {largest_slope} has the largest slope,
     there are {sum(faster_react)} subject that show an increacment in reaction time,
     there are {sum(neg_inetr_pos_slope)} subjects with negative intercept and positive slope,
     there are {pos_inetr_neg_slope} subjects with positive intercept and negative slope,
     for model 2")
```

```
## subject 10 has the largest intercept,
## there are 7 (39%) subject with negative intercept,
## subject 1 has the largest slope,
## there are 17 subject that show an increacment in reaction time,
## there are 6 subjects with negative intercept and positive slope,
## there are 0 subjects with positive intercept and negative slope,
## for model 2
```

there is no really negative intercepts i assumed that be negative intercepts you meant in comparison to the population

Q3

שאלה 3

בשיעור הגדרנו RD (risk difference), RR (risk ratio) ו- OR (odds ratio).

הוכיחו כי

א. אם RD = 0, אז בהכרח מתקיים כי RR = 1 וגם OR = 1.

ב. אם RD > 0, אז בהכרח מתקיים כי RR > 1 וגם OR > 1.

a

$$Rd = Pr(Y=1|X=1) - Pr(Y=1|X=0)$$

$$Rd = 0 \iff Pr(Y=1|X=1) = Pr(Y=1|X=0) \iff RR = \frac{Pr(Y=1|X=1)}{Pr(Y=1|X=0)} = 1$$

$$OR = \frac{\frac{Pr(Y=1|X=1)}{Pr(Y=0|X=1)}}{\frac{Pr(Y=1|X=0)}{Pr(Y=0|X=0)}} = \frac{Pr(Y=1|X=1)Pr(Y=0|X=0)}{Pr(Y=0|X=1)Pr(Y=0|X=1)} = RR * \frac{Pr(Y=0|X=0)}{Pr(Y=0|X=1)}$$

$$\frac{Pr(Y=0|X=0)}{Pr(Y=0|X=1)} = \frac{1-Pr(Y=1|X=0)}{1-Pr(Y=1|X=1)}$$

$$Rd = 0 \iff RR\frac{1-Pr(Y=1|X=0)}{1-Pr(Y=1|X=1)} = RR\frac{1-Pr(Y=1|X=1)}{1-Pr(Y=1|X=1)} \iff OR = 1$$

b

$$Rd = Pr(Y=1|X=1) - Pr(Y=1|X=0)$$

$$Rd > 0 \iff Pr(Y=1|X=1) > Pr(Y=1|X=0) \iff RR = \frac{Pr(Y=1|X=1)}{Pr(Y=1|X=0)} > 1$$

$$OR = \frac{\frac{Pr(Y=1|X=1)}{Pr(Y=0|X=1)}}{\frac{Pr(Y=1|X=0)}{Pr(Y=0|X=0)}} = \frac{Pr(Y=1|X=1)Pr(Y=0|X=0)}{Pr(Y=0|X=1)Pr(Y=0|X=1)} = RR * \frac{Pr(Y=0|X=0)}{Pr(Y=0|X=1)}$$

$$Rd > 0 \iff RR\frac{1-Pr(Y=1|X=0)}{1-Pr(Y=1|X=1)} > \frac{1-Pr(Y=1|X=1)}{1-Pr(Y=1|X=1)} \iff OR > 1$$

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

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

א. חשבו אומדים נקודתיים ל RD,RR,OR עבור מגדר כמשתנה מסביר ותמותה בעקבות אוטם ראשון בשריר הלב כמשתנה מוסבר.

ב. הסבירו במילים מה המשמעות של כל אחד מהאומדים שהתקבל.

ג. הסבירו עבור כל אחד מבין RD,RR,OR מדוע הוא מתאים - או לא מתאים - לניתוח נתונים אלו.

a

```
MI <- read.csv("MI_PracticeDataset.csv") %>%
  dplyr::select(Sex,CVDeath_2012)
pr_death_man <-  mean(MI[MI$Sex==1,]$CVDeath_2012)
pr_death_woman <-  mean(MI[MI$Sex==2,]$CVDeath_2012)
RD <- pr_death_man - pr_death_woman
RR <- pr_death_man/pr_death_woman
OR <- (pr_death_man/(1-pr_death_man))/(pr_death_woman/(1-pr_death_woman))
glue("RD:{RD}
     RR:{RR}
     OR:{OR}")
```

```
## RD:-0.0722290252999595
## RR:0.725339627477785
## OR:0.660600221717785
```

b

the RD estimates suggests that a man has 7% lower chance of dying from myocardial infection

the RR estimates suggests that a man chance of dying from myocardial infection is 72.5% of the chance woman has

the OR estimates suggests that a man chance of dying from myocardial infection compared to not getting an infection is 66% of the chance woman has

c

RD is good to use because it tells us that man has 7% less chance of dying from myocardial infection

RR by its own does not give us much information it tells us that a man chance of dying from myocardial infection is 72.5% of the chance woman has, but if woman chance is low then this does not provide us important information

OR is good to use because it tell us that man are less prone to dying from myocardial infection