ST3243 SEMESTER 1 2016/17

Project Assignment

Section I

A: Data Description

Abstract

Outcome: To predict probability of survival of patients after discharged from ICU from hospital

Participants: 400 re-sampled subjects from a larger study on survival of patients admission into ICU

Design: Logistic Regression

Main outcome measure: Vital status at hospital discharge: STA. Dichotomous variable.

0 = Lived and 1 = Died

Key determinant: Patients' age, AGE. Continuous variable. For this data set, is 16 to 92 years old.

R code for data input:

```
data <- read.delim("D:/NUS-sync/ST3243 Statistical Methods In Epidemiology/Project/Questions and Datasets/icu.data.s10.txt", header=FALSE)
sta<-data$V2
age<-data$V3
```

Section B

Question 1A

Logistic regression model for STA on Age :

$$p = \frac{1}{(1 + e^{-(\alpha+\beta age)})}$$

Logistic regression model for STA on Age with logit transformation:

$$\text{logit(p)} = \ln\left(\frac{p}{1-p}\right) = \alpha + \beta x$$

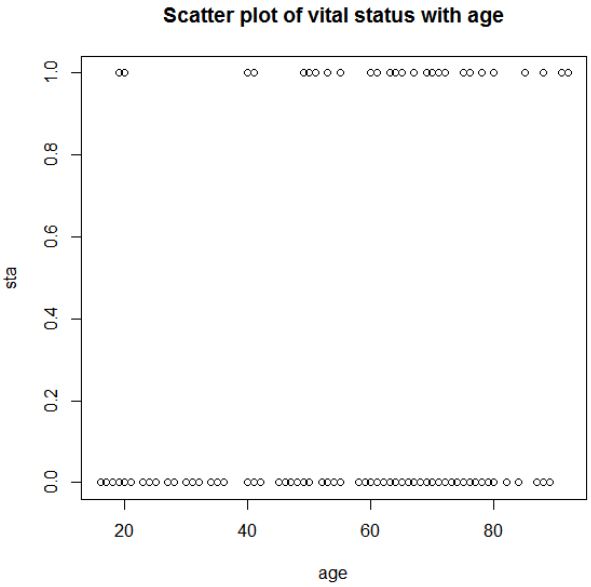
Outcome variable, STA, is a dichotomous variable or binary outcome and non-negative. Hence, this leads us to consider logistic regression model as opposed to the usual linear.

Question 1B

R code for plotting STA vs Age of the re-sampled 400 subjects

```
plot(age,sta,xlab="age",ylab="sta",main="Scatter plot of vital status with age")
```

Scatter plot of STA versus Age



Question 1C

Re-coding of the age of subjects into 8 intervals

```
attach(data)
data$ageinterval[age>=15 & age<=24]="[15,24]"
data$ageinterval[age>=25 & age<=34]="[25,34]"
data$ageinterval[age>=35 & age<=44]="[35,44]"
data$ageinterval[age>=45 & age<=54]="[45,54]"
data$ageinterval[age>=55 & age<=64]="[55,64]"
data$ageinterval[age>=65 & age<=74]="[65,74]"
data$ageinterval[age>=75 & age<=84]="[75,84]"
data$ageinterval[age>=85 & age<=94]="[85,94]"
detach(data)
```

Using the intervals [15, 24], [25, 34], [35, 44], [45, 54], [55, 64], [65, 74], [75, 84], [85, 94] for AGE, and computed the STA mean of subjects within each AGE interval

```
stameans=tapply(sta,data$ageinterval,mean)
```

```
> stameans
   [15,24]    [25,34]    [35,44]    [45,54]    [55,64]    [65,74]    [75,84]    [85,94]
0.1276596 0.0000000 0.2000000 0.2745098 0.2000000 0.1538462 0.2280702 0.5714286
```

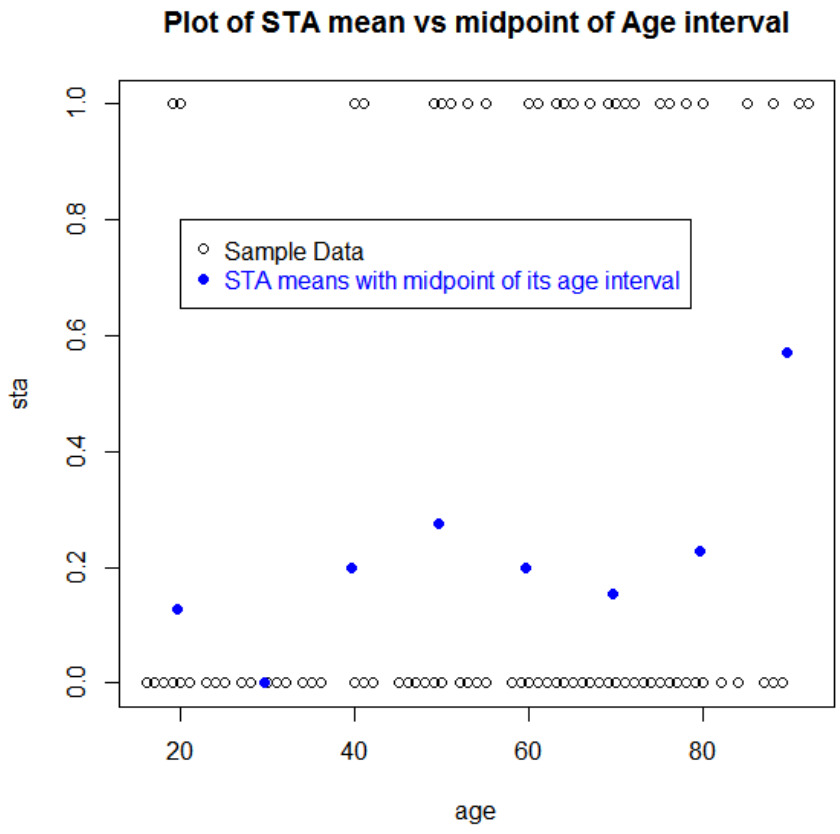
Tabulating the output above into a table format rounding off to 4 significant figures

Age interval	[15,24]	[25,34]	[35,44]	[45,54]	[55,64]	[65,74]	[75,84]	[85,94]
STA mean	0.1277	0	0.2000	0.2745	0.2000	0.1538	0.2281	0.5714

Table 1: STA mean within each age interval

Plotting means of STA of subjects in each age intervals vs midpoint of their respective interval, over the same plot from question 1B

```
midpoint=c(median(15:24),median(25:34),median(35:44),median(45:54),median(55:64),median(65:74),median(75:84),median(85:94))
plot(age,sta,main="Plot of STA mean vs midpoint of Age interval",pch=1)
points(midpoint,stameans,col="blue",pch=16)
legend(20，0.8，c("Sample Data","STA means with midpoint of its age interval"),text.col=c("black","blue"),pch=c(1,16),col=c("black","blue"))
```



Question 1D

R code for logistic regression model in question 1A based on ungrouped n=400 data

```
logit=glm(sta~age, family=binomial(link="logit"))
summary(logit)
```

```
Call:
glm(formula = sta ~ age, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8757  -0.7312  -0.6511  -0.4614   2.1330

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.54787    0.45155  -5.643 1.68e-08 ***
age          0.02008    0.00705   2.848   0.0044 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 396.96  on 398  degrees of freedom
AIC: 400.96

Number of Fisher Scoring iterations: 4
```

From R output:

Intercept ($\hat{\alpha}$) = -2.54787, coefficient of Age ($\hat{\beta}$) = 0.02008

Fitted values for the estimated logistic regression model with logit transformation:

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\alpha} + \hat{\beta} age = -2.54787 + 0.02008(age)$$

$$\approx -2.5479 + 0.02008 \text{ (age) (in 4 sf.)}$$

Fitted values for the estimated logistic regression model:

$$\hat{p} = \frac{1}{\left(1+e^{-(\hat{\alpha}+\hat{\beta} age)}\right)} = \frac{1}{(1+e^{-(-2.54787+0.02008\, age)})} \approx \frac{1}{(1+e^{2.548-0.02008\, age})} \quad \text{(in 4s.f)}$$

R code for fitted values

```
fitted=1/(1+ exp(-coef(logit)[1]-coef(logit)[2]*age))
```

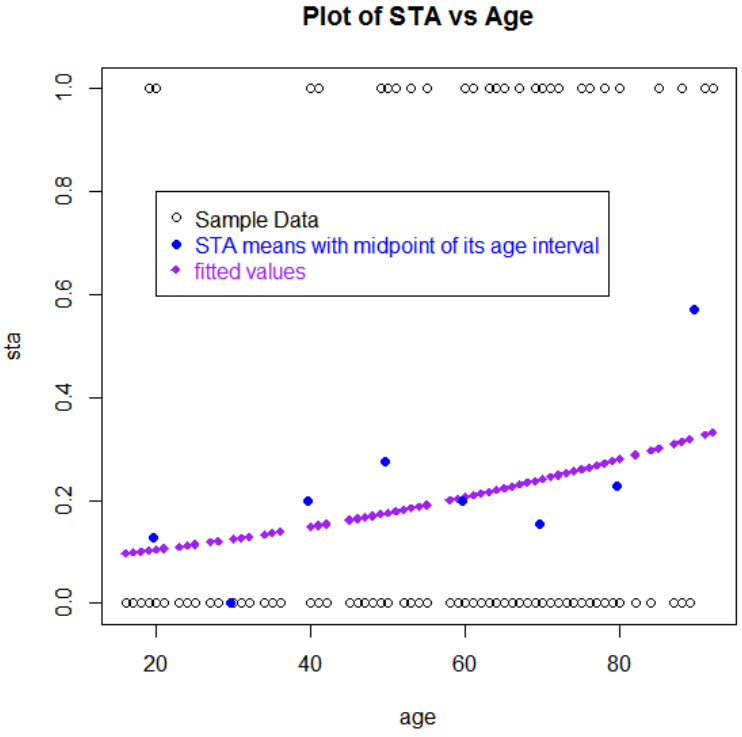
First few outputs:

```
> fitted
  [1] 0.14871227 0.26857946 0.22746122 0.21365796 0.22048213 0.16739694
  [7] 0.24188192 0.23100871 0.26465357 0.22048213 0.22746122 0.17596116
 [13] 0.21365796 0.21030405 0.16739694 0.20371257 0.26076461 0.28875250
 [19] 0.13644594 0.14871227 0.19099418 0.31410848 0.23459473 0.23821918
 [25] 0.22395234 0.26076461 0.10097316 0.21030405 0.16187462 0.24932173
 [31] 0.18486625 0.16187462 0.20371257 0.24558283 0.18486625 0.21030405
 [37] 0.10467732 0.10281047 0.10281047 0.22746122 0.24932173 0.24188192
 [43] 0.12725605 0.24558283 0.23459473 0.16739694 0.19099418 0.16187462
 [49] 0.22048213 0.22048213 0.29706905 0.22048213 0.30127878 0.10467732
 [55] 0.17306876 0.20371257 0.29706905 0.23459473 0.27254200 0.13644594
 [61] 0.14871227 0.10467732 0.23821918 0.24188192 0.26857946 0.22048213
 [67] 0.17596116 0.23459473 0.11045843 0.30127878 0.31410848 0.10467732
 [73] 0.23459473 0.24932173 0.19099418 0.16461719 0.15127208 0.28057584
 [79] 0.22048213 0.31410848 0.23821918 0.10097316 0.27254200 0.16739694
 [85] 0.20698892 0.31845028 0.23100871 0.31410848 0.17306876 0.09738581
 [91] 0.28057584 0.23100871 0.22395234 0.19099418 0.10657405 0.21030405
 [97] 0.21705066 0.16739694 0.09738581 0.18486625 0.23821918 0.14871227
[103] 0.20047497 0.26076461 0.19099418 0.22048213 0.22048213 0.13409737
[109] 0.15386800 0.24188192 0.31410848 0.23100871 0.12071510 0.20698892
[115] 0.21030405 0.22746122 0.24188192 0.23100871 0.19099418 0.20698892
[121] 0.21030405 0.20698892 0.18486625 0.13882904 0.24932173 0.17306876
[127] 0.21030405 0.26076461 0.11045843 0.20371257 0.25309847 0.10281047
[133] 0.11446626 0.22746122 0.12950272 0.22048213 0.20371257 0.19099418
```

R code for plot of fitted values of the logistic regression on the scatter plot in question 1B and 1C

```
plot(age,sta,main="Plot of STA vs Age",pch=1)
points(age,fitted,pch=18,col="purple")
points(midpoint,stameans,col="blue",pch=16)
legend(20，0.8，c("Sample Data","STA means with midpoint of its age interval","fitted values"),text.col=c("black","blue","purple"),col=c("black","blue","purple"),pch=c(1,16,18))
```

Plot of equation for the fitted values on the axes used in the scatter plot in Questions 1B and 1C



Question 1E

Summarizing the results in plot from question 1B

From the scatter plot of STA vs age, which is a dichotomous variable against a continuous variable, at first sight seems somewhat uninformative. However it gives some vague idea about the data we are dealing with.

Overall the scatter plot shows a higher number of total points, for both STA =0 or 1, between age groups 55 to 80 years old. It can be simply seen by imagining a middle line cutting through at age 55 years old disregarding which STA the points fall in; one can observe that there is higher total number of points in the right region (older people). This suggests that more old people were admitted into ICU to begin with.

Looking at Vital Status = 1 (died), the age distribution looks left skewed: between age of 60 to 80 years old having more clustered number of points. The interpretation is that among those who died upon admission to ICU, there is a higher proportion of old people compared to younger people. However, we deduced earlier that more old people were admitted into ICU among the 400 subjects to begin with. Hence, the proportion of old people who die upon admission in to ICU may actually be similar as that of the younger people.

Similarly like Vital Status = 0 (lived), there is slightly more clustering between aged of 55 to 80 years old compared to 15 to 55 years old for Vital Status = 0 (lived) which suggests that looking at those who lived upon admission into ICU, there is a higher proportion of old people compared to younger people. However, that is a weak statement since there are seemingly more old people admitted into ICU to begin with. More information is needed to make a good deduction of whether it is more likely to observe old people if one were to die upon admission into ICU.

Summarizing the results in plot from question 1C

From the scatter plot we can observe more things than from question 1B , we can observe that STA mean hovers around 0.2 for age group 15 till 85 years old. The last age group 85 to 94 years old seems to be an outlier. Hence, it seems to suggest that age is not a confounder. By also using STA mean instead of STA, we eliminate the effect of different sampling proportion for different age group influence on STA.

Summarizing the results in plot from question 1D

Plot in question 1D is a fit for the model for the ungrouped data n=400. We can observe that there is a clear positive association. As age increases, STA increases. However, it is possible that the outlier mentioned earlier in the last age group causing this association, thus resulting in a poor fit. Looking at age group below 85 years old, the increase in STA with age seems gradual. This plot gives a rough idea that age may be a confounder on the ungrouped data.

Question 1F

Using the results of the output for Question 1D for logistic regression of STA vs Age for n=400, assess the significance of the slope coefficient for Age

Recalling the results

```
Call:
glm(formula = sta ~ age, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8757  -0.7312  -0.6511  -0.4614   2.1330

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.54787    0.45155  -5.643 1.68e-08 ***
age          0.02008    0.00705   2.848   0.0044 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 396.96  on 398  degrees of freedom
AIC: 400.96

Number of Fisher Scoring iterations: 4
```

The slope coefficient of the logistic regression for question 1D is the coefficient of Age, which is $\hat{\beta}$ = 0.02008, with Standard Error = 0.00705, Z value= 2.848 and p-value of 0.0044. Hence, we can deduce that age is statistically significant at 5% level since 0.0044<0.05. This meant that age is a good predictor for STA. In addition, age is has positive relationship with STA. And for every unit increase in age, log odds ratio of STA increases by a factor of 0.02008. This translates to every unit increase in age, odds ratio of STA increases by a factor of $e^{0.02008}$ = 1.02028 ≈ 1.020 (4 s.f). This meant that the odds of having STA=1 is 1.020 times odds of having STA =0 for every unit increase in age, which is not very significant. However, the coefficient of age is still statistically significant at 5% level.

Question 1G

Using the results from Question 1D

For slope term for the logistic regression of STA vs Age,

Coefficient of Age is $\hat{\beta}$ = 0.02008, with Standard Error = 0.00705

For intercept for the logistic regression of STA vs Age,

$\hat{\alpha}$ = -2.54787, with Standard Error = 0.45155

Computing the 95% Confidence Intervals (CI) for the slope and intercept term

For slope term for the logistic regression of STA vs Age,

95% Confidence Interval = $e^{0.02008 \pm 1.96\ (0.00705)}$ = (1.00628, 1.03447) ≈ (1.006, 1.034) (4 s.f)

For intercept term for the logistic regression of STA vs Age,

95% Confidence Interval = $e^{-2.54787 \pm 1.96\ (0.45155)}$ = (0.0322928, 0.189601) ≈ (0.03229, 0.1896) (4 s.f)

Write a sentence interpreting the CI for the slope

Given the 95% Confidence Interval of the slope (1.006, 1.034) does not contain 1, we can conclude that age is a statistically significant predictor and there is positive association between age and vital status.

R Code to check answer

```
upperlimit1=exp(summary(logit)$coefficients[2,1]+1.96*summary(logit)$coefficients[2,2])
lowerlimit1=exp(summary(logit)$coefficients[2,1]-1.96*summary(logit)$coefficients[2,2])
confidenceintervalforslope=c(lowerlimit1,upperlimit1)
upperlimit2=exp(summary(logit)$coefficients[1,1]+1.96*summary(logit)$coefficients[1,2])
lowerlimit2=exp(summary(logit)$coefficients[1,1]-1.96*summary(logit)$coefficients[1,2])
confidenceintervalforintercept=c(lowerlimit2,upperlimit2)
> confidenceintervalforslope
[1] 1.006279 1.034478
> confidenceintervalforintercept
[1] 0.03229325 0.18960046
```

OR exp(confint(logit))

```
Waiting for profiling to be done...
                2.5 %     97.5 %
(Intercept) 0.03077985 0.1818418
age         1.00669380 1.0349965
```

Question 1H

Compute the logit and estimated logistic probability for a 60-year old subject. Compute a 95% CI for the logit and estimated logistic probability.

The logit for a 60 years old subject is

$$\text{logit}(\hat{p}_i) = \hat{\alpha} + \hat{\beta} age = -2.54787 + 0.02008(60) = -1.34307 \approx -1.343 \text{ (4 s.f)}$$

The estimated probability for a 60 years old subject is

$$\hat{p} = \frac{1}{\left(1+e^{-(\hat{\alpha}+\hat{\beta} age)}\right)} = \frac{1}{1+e^{2.54787-0.02008\,(60)}} = 0.207005 \approx 0.2070 \text{ (4 s.f)}$$

R Code to check answer

```
logitfor60yearsold=1/(1+exp(-summary(logit)$coefficients[1,1]-60*summary(logit)$coefficients[2,1]))
```

```
> logitfor60yearsold
[1] 0.2069889
```

To compute the 95% CI for a 60 years old subject, we need to compute the variance of the logit.

R Code for variance-covariance matrix

```
vcov(logit)
```

```
             (Intercept)           age
(Intercept)  0.203893865 -3.058764e-03
age         -0.003058764  4.970756e-05
```

Presenting the variance-covariance matrix in a table

	Intercept (α)	Age (β)
Intercept (α)	0.203893865	-0.003058764
Age (β)	-0.003058764	0.00004970756

Table 2: Variance-covariance matrix for logistic regression model

$$var[\text{logit}(\hat{p}_i)] = var[\hat{\alpha} + \hat{\beta}.(60)] = var(\hat{\alpha}) + 60^2\, var(\hat{\beta}) + 2\,(60)\, cov(\hat{\alpha} + \hat{\beta})$$

$$= 0.203893865 + 60^2\,(0.00004970756) + 120\,(-0.003058764)$$

$$= 0.015789401$$

$$\approx 0.01579 \text{ (4 s.f)}$$

$$SE[\text{logit}(\hat{p}_i)] = \sqrt{var[\text{logit}(\hat{p}_i)]} = \sqrt{0.015789401} = 0.125655 \approx 0.1257 \text{ (4 s.f)}$$

Compute the 95% CI for a 60 years old subject

95% Confidence Interval for the logit = $\text{logit}(\hat{p}_i) \pm 1.96 \,.\, se[\text{logit}(\hat{p}_i)]$

= -1.34307 $\pm$ 1.96 (0.125655)

= (-1.58935 ,-1.09678)

$\approx$ (-1.589, -1.097) (4 s.f)

95% Confidence Interval for the estimated probability = $\left(\frac{1}{1+e^{-(-1.58935)}}, \frac{1}{1+e^{-(-1.09678)}}\right)$

$=$ (0.169475, 0.250343)

$\approx$ (0.1695, 0.2503) (4 s.f)

Write a sentence or two interpreting the estimated probability and its CI.

We are 95% confident that the probability of a 60 years old subject having STA falls within (0.1695, 0.2503) and its estimated STA is 0.2070. Since it is closer to 0 than 1, there is a higher probability that a 60 years old subject will live than die.

Question 1 I

Using R, obtain the estimated logit and its standard error for each subject in the ICU study.

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\alpha} + \hat{\beta} age$$

```
logitmodel=coef(logit)[1]+coef(logit)[2]*age
```

The estimated logit is

```
> logitmodel
  [1] -1.7447368 -1.0018418 -1.2227025 -1.3030154 -1.2628589 -1.6041891
  [7] -1.1423895 -1.2026242 -1.0219200 -1.2628589 -1.2227025 -1.5439544
 [13] -1.3030154 -1.3230937 -1.6041891 -1.3632502 -1.0419983 -0.9014505
 [19] -1.8451280 -1.7447368 -1.4435631 -0.7809811 -1.1825460 -1.1624677
 [25] -1.2427807 -1.0419983 -2.1864582 -1.3230937 -1.6443456 -1.1022330
 [31] -1.4837196 -1.6443456 -1.3632502 -1.1223112 -1.4837196 -1.3230937
 [37] -2.1463017 -2.1663799 -2.1663799 -1.2227025 -1.1022330 -1.1423895
 [43] -1.9254410 -1.1223112 -1.1825460 -1.6041891 -1.4435631 -1.6443456
 [49] -1.2628589 -1.2628589 -0.8612941 -1.2628589 -0.8412158 -2.1463017
 [55] -1.5640326 -1.3632502 -0.8612941 -1.1825460 -0.9817635 -1.8451280
 [61] -1.7447368 -2.1463017 -1.1624677 -1.1423895 -1.0018418 -1.2628589
 [67] -1.5439544 -1.1825460 -2.0860670 -0.8412158 -0.7809811 -2.1463017
 [73] -1.1825460 -1.1022330 -1.4435631 -1.6242673 -1.7246586 -0.9416070
 [79] -1.2628589 -0.7809811 -1.1624677 -2.1864582 -0.9817635 -1.6041891
 [85] -1.3431719 -0.7609028 -1.2026242 -0.7809811 -1.5640326 -2.2266147
 [91] -0.9416070 -1.2026242 -1.2427807 -1.4435631 -2.1262235 -1.3230937
 [97] -1.2829372 -1.6041891 -2.2266147 -1.4837196 -1.1624677 -1.7447368
[103] -1.3833284 -1.0419983 -1.4435631 -1.2628589 -1.2628589 -1.8652063
[109] -1.7045803 -1.1423895 -0.7809811 -1.2026242 -1.9856757 -1.3431719
[115] -1.3230937 -1.2227025 -1.1423895 -1.2026242 -1.4435631 -1.3431719
[121] -1.3230937 -1.3431719 -1.4837196 -1.8250498 -1.1022330 -1.5640326
[127] -1.3230937 -1.0419983 -2.0860670 -1.3632502 -1.0821547 -2.1663799
[133] -2.0459105 -1.2227025 -1.9053628 -1.2628589 -1.3632502 -1.4435631
[139] -1.7246586 -1.1624677 -0.9014505 -2.0057540 -0.9014505 -1.3431719
```

$$var[\text{logit}(\hat{p}_i)] = var[\hat{\alpha} + \hat{\beta}.(age)] = var(\hat{\alpha}) + age^2\, var(\hat{\beta}) + 2\,(age)\, cov(\hat{\alpha} + \hat{\beta})$$

$$SE[\text{logit}(\hat{p}_i)] = \sqrt{var[\text{logit}(\hat{p}_i)]}$$

```
varcov=vcov(logit)
```

```
> varcov
             (Intercept)           age
(Intercept)  0.203893865 -3.058764e-03
age         -0.003058764  4.970756e-05
```

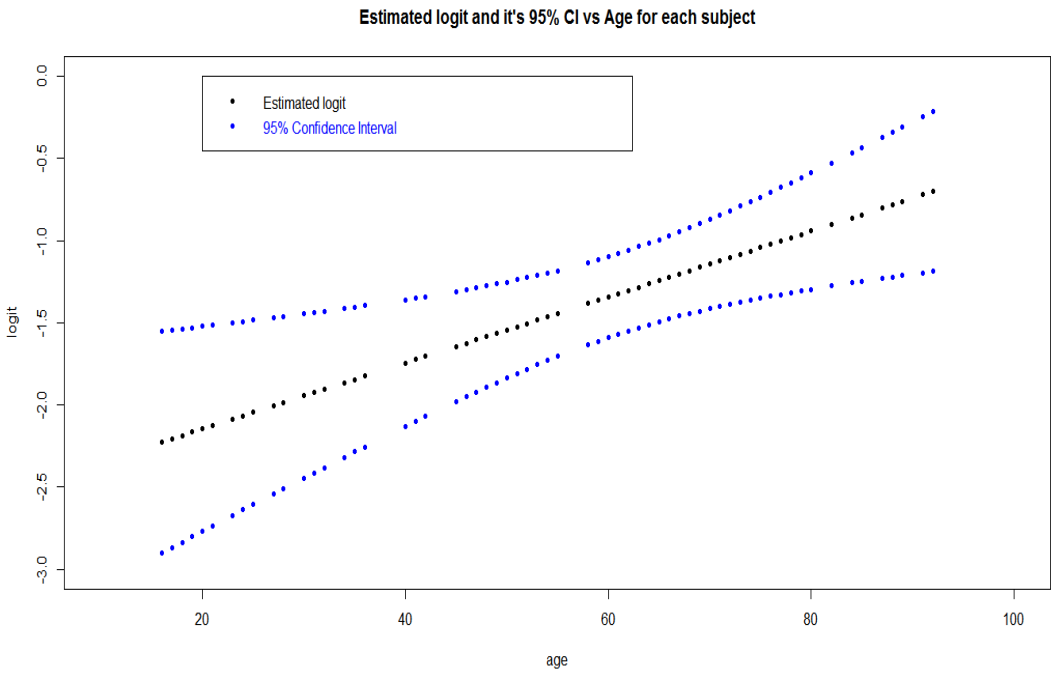
```
varlogit= function(age) return (varlogit= varcov[1,1] + (age^2)* (varcov[2,2])+ 2*age*varcov[2,1])
selogit=rep(0,400)
for (k in 1:400){
selogit[k]=sqrt(varlogit(age[k]))
}
```

The standard error for each subject in the ICU study

```
> selogit
  [1] 0.1967862 0.1660129 0.1290857 0.1252316 0.1263891 0.1617839 0.1386864
  [8] 0.1309836 0.1614700 0.1263891 0.1290857 0.1492861 0.1252316 0.1252456
 [15] 0.1617839 0.1264583 0.1571122 0.1910240 0.2251047 0.1967862 0.1333985
 [22] 0.2246925 0.1332279 0.1358017 0.1275498 0.1571122 0.3314870 0.1252456
 [29] 0.1710640 0.1453128 0.1389006 0.1710640 0.1264583 0.1418631 0.1389006
 [36] 0.1252456 0.3184750 0.3249696 0.3249696 0.1290857 0.1453128 0.1386864
 [43] 0.2490370 0.1418631 0.1332279 0.1617839 0.1333985 0.1710640 0.1263891
 [50] 0.1263891 0.2018861 0.1263891 0.2074634 0.3184750 0.1532410 0.1264583
 [57] 0.2018861 0.1332279 0.1707262 0.2251047 0.1967862 0.3184750 0.1358017
 [64] 0.1386864 0.1660129 0.1263891 0.1492861 0.1332279 0.2991421 0.2074634
 [71] 0.2246925 0.3184750 0.1332279 0.1453128 0.1333985 0.1663393 0.1913991
 [78] 0.1806100 0.1263891 0.2246925 0.1358017 0.3314870 0.1707262 0.1617839
 [85] 0.1256558 0.2305806 0.1309836 0.2246925 0.1532410 0.3445846 0.1806100
 [92] 0.1309836 0.1275498 0.1333985 0.3120045 0.1252456 0.1256140 0.1617839
 [99] 0.3445846 0.1389006 0.1358017 0.1967862 0.1276458 0.1571122 0.1333985
[106] 0.1263891 0.1263891 0.2309975 0.1861231 0.1386864 0.2246925 0.1309836
[113] 0.2675328 0.1256558 0.1252456 0.1290857 0.1386864 0.1309836 0.1333985
[120] 0.1256558 0.1252456 0.1256558 0.1389006 0.2192803 0.1453128 0.1532410
[127] 0.1252456 0.1571122 0.2991421 0.1264583 0.1490163 0.3249696 0.2863964
[134] 0.1290857 0.2429681 0.1263891 0.1264583 0.1333985 0.1913991 0.1358017
[141] 0.1910240 0.2737835 0.1910240 0.1256558 0.1571122 0.1710640 0.1710640
[148] 0.3380257 0.1418631 0.1571122 0.1333985 0.1614700 0.1256558 0.1571122
[155] 0.3380257 0.1617839 0.1309836 0.1256558 0.1358017 0.1252456 0.1453128
[162] 0.1275498 0.1389006 0.3249696 0.1913991 0.1359947 0.1617839 0.2246925
[169] 0.1420972 0.3184750 0.2429681 0.1332279 0.1252316 0.1386864 0.1913991
```

Graph the estimated logit and the point-wise 95% confidence limits versus AGE for each subject.

```
plot(age,logitmodel,main="Estimated logit and it's 95% CI vs Age for each subject",ylab="logit",
xlab="age",pch=20,col="black",xlim=c(10,100),ylim=c(-3,0))
points(age,logitmodel+1.96*selogit,col="blue",pch=20)
points(age,logitmodel-1.96*selogit,col="blue",pch=20)
legend(20，0，c("Estimated logit","95% Confidence Interval"),text.col=c("black","blue"),
pch=c(20,20),col=c("black","blue"))
```



Explain in words the similarities and differences between the appearance of this graph and a graph of a fitted linear regression model and its point-wise 95% confidence bands.

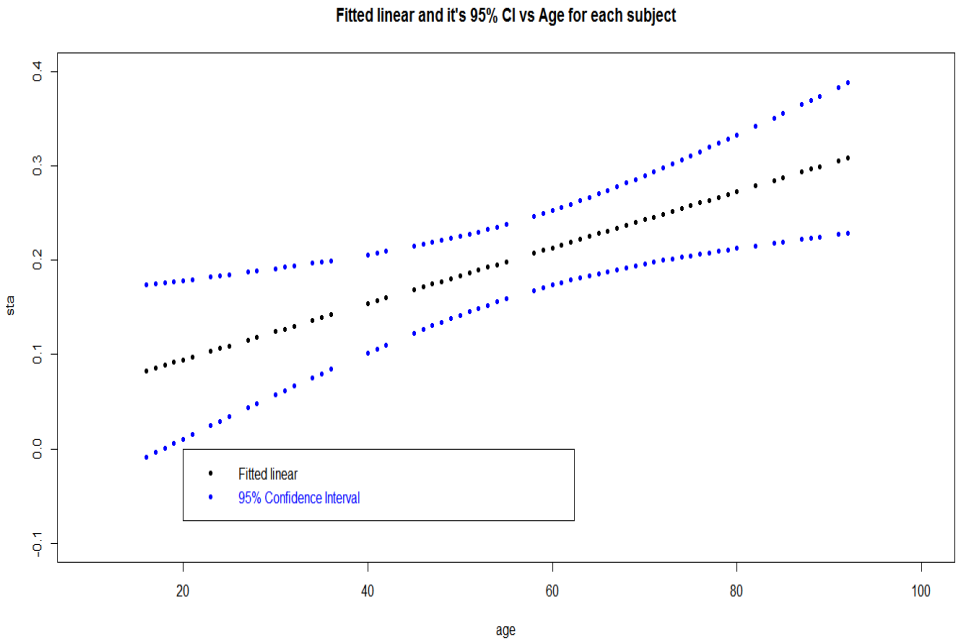
The shape of graph of fitted linear regression will be similar to that of the logit graph. This is because both the logit and fitted linear are in a linear form (Y=a+bX): the fit will be a straight line with coefficient and intercept from the fit. The confidence interval surrounding the fit will be of similar shape. Logit is like log linear from regressions with coefficients of exponents form.

The difference is the range of y values which are the fitted estimates and that logit is from regression with coefficients of exponents form.

R code to fit linear regression model to illustrate (extra)

```
fitlinearmodel=lm(sta~age)
fitlinear=coef(fitlinearmodel)[1]+coef(fitlinearmodel)[2]*age
varcov1=vcov(fitlinearmodel)
varlinear= function(age) return (varlinear= varcov1[1,1] + (age^2)* (varcov1[2,2])+
2*age*varcov1[2,1])
selinear=rep(0,400)
for (k in 1:400){
selinear[k]=sqrt(varlinear(age[k]))
}
plot(age,fitlinear,main="Fitted linear and it's 95% CI vs Age for each subject",ylab=" sta",xlab="age",
pch=20,col="black",xlim=c(10,100),ylim=c(-0.1,0.4))
points(age,fitlinear+1.96*selinear,col="blue",pch=20)
points(age,fitlinear-1.96*selinear,col="blue",pch=20)
```

```
legend(20，0，c("Fitted linear","95% Confidence Interval"),text.col=c("black","blue"),
pch=c(20,20),col=c("black","blue"))
```



Section B

Question 2 A

Demonstrate that the value of the log odds ratio obtained from the cross-tabulation of STA by CPR is identical to the estimated slope coefficient from the logistic regression of STA on CPR.

To compute the odds ratio from the cross-tabulation

R code to get cross table for STA vs CPR

```
cpr<-data$V10
xtabs(~cpr+sta,data=data)
```

```
    sta
cpr   0   1
  0 302  70
  1  16  12
```

Putting the cross tabulation into a table

	STA = 1	STA = 0
CPR = 1	12	16
CPR = 0	70	302

Table 3: Cross tabulation of STA vs CPR

Using the cross tabulation table 3

$$\widehat{\text{OR}}(\text{CPR} = 1 \text{ vs CPR} = 0) = \frac{12*302}{16*70} = 3.23571 \approx 3.236 \text{ (4 s.f)}$$

$$ln\left[\widehat{\text{OR}}(\text{CPR} = 1 \text{ vs CPR} = 0)\right] = \ln(3.23571) = 1.17424 \approx 1.174 \text{ (4 s.f)}$$

The log odds ratio from the cross-tabulation is *1.174* (4 s.f)

To compute the slope coefficient from logistic regression of STA on CPR

R code to get slope coefficient $(\hat{\beta}_1)$ from logistic regression of STA on CPR

```
cprlogit=glm(sta~cpr, family=binomial(link="logit"))
slopeofcpr=summary(cprlogit)$coefficients[2,1]
```

```
> slopeofcpr
[1] 1.17425
```

The estimated slope coefficient from the logistic regression is 1.17425 ≈ *1.174* (4 s.f)

Hence, we have shown that value of the log-odds ratio obtained from the cross-tabulation of STA by CPR is identical to the estimated slope coefficient from the logistic regression of STA on CPR.

Verify that the estimated standard error of the estimated slope coefficient for CPR obtained from the R or other statistical package is identical to the square root of the sum of the inverse of the cells frequencies from the cross-tabulation of STA by CPR.

R code to get the estimated standard error of the slope coefficient $(\hat{\beta}_1)$ for CPR

```
secpr=summary(cprlogit)$coefficients[2,2]
```

```
> secpr
[1] 0.4042651
```

The estimated standard error of the estimated slope coefficient for CPR from the R is approximately 0.4043 (4 s.f)

Recalling table 3 earlier,

	STA = 1	STA = 0
CPR = 1	12	16
CPR = 0	70	302

Table 3: Cross tabulation of STA vs CPR

$$\widehat{SE}[\text{logit}(\hat{p}_i)] = \widehat{SE}\left[\, ln\left[\, \widehat{OR}(CPR = 1 \; vs \; CPR = 0)\right]\right.$$

$$\approx \sqrt{\frac{1}{12} + \frac{1}{16} + \frac{1}{70} + \frac{1}{302}} = 0.404265 \approx 0.4043 \text{ (4 s.f)}$$

The square root of the sum of the inverse of the cells frequencies from the cross-tabulation of STA by CPR is 0.4043 (4 s.f)

Use either set of computations to obtain 95% CI for the odds ratio

95% Confidence interval for the log odds ratio for CPR = 1 vs. CPR = 0

$$ln[\widehat{OR}(CPR = 1\ vs.\ CPR = 0)] \pm 1.96 * \widehat{SE}(ln[\widehat{OR}(CPR = 1\ vs.\ CPR = 0)])$$

95% Confidence interval for the odds ratio for CPR = 1 vs. CPR = 0

$$e^{ln[\widehat{OR}(CPR=1\ vs.\ CPR=0)]\pm 1.96*\widehat{SE}(ln[\widehat{OR}(CPR=1\ vs.\ CPR=0)])}$$

Using the cross tabulation results earlier,

The 95% confidence interval for log odds ratio is $1.17425 \pm 1.96\ (0.404265)$

= (0.381891, 1.96661)

≈ (0.3819, 1.967) (4 s.f)

The 95% confidence interval for odds ratio is = $e^{1.17425 \pm 1.96\ (0.404265)}$

= (1.46505, 7.14641)

≈ (1.465, 7.146) (4 s.f)

What aspect concerning the coding of the variable CPR makes the calculations for the two methods equivalent?

Variable CPR is dichotomous.

Question 2 B

For this problem only, variable CPR refers to CPR with no =4 and yes =2

Use a data transformation statement to recode the variable CPR as follows: 4=no and 2=yes

Using R to recoded CPR with yes=2 and no=4

```
recodedcpr=data$recodedcpr
recodedcpr=ifelse(cpr == 0, 4, 2)
```

```
> recodedcpr
  [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 2 4 4 4
 [38] 4 4 4 4 4 4 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4
 [75] 4 4 4 4 4 4 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 2 4 4 4 4 4 4 4 4 4
[112] 4 4 4 4 4 4 4 4 2 4 4 2 4 4 2 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 4 4 4 4 4 4
[149] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 4 4 4 2 2 4 4
[186] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 4 4
[223] 4 4 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 4 4 4 4 4 4
[260] 4 4 4 4 4 4 4 4 4 4 4 4 2 4 2 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 2 4 4 4 4 4
[297] 2 4 4 4 4 2 2 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4
[334] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
[371] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 2 4 4 4 4 2 4 4 4 4 4 4 4
> cpr
  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0
 [38] 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
 [75] 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0
[112] 0 0 0 0 0 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
[149] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0
[186] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0
[223] 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
[260] 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0
[297] 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
[334] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[371] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0
```

Perform the logistic regression of STA on CPR (recoded)

Using R to compute logistic regression of STA vs CPR (recoded)

```
recodedcprlogit=glm(sta~recodedcpr, family=binomial(link="logit"))
summary(recodedcprlogit)$coefficients[2,1]
```

```
> summary(recodedcprlogit)

Call:
glm(formula = sta ~ recodedcpr, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0579  -0.6457  -0.6457  -0.6457   1.8278

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8866     0.7752   1.144  0.25276
recodedcpr   -0.5871     0.2021  -2.905  0.00368 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 398.01  on 398  degrees of freedom
AIC: 402.01

Number of Fisher Scoring iterations: 4


> summary(recodedcprlogit)$coefficients[2,1]
[1] -0.5871249
```

The coefficient of the slope recoded CPR is -0.5871249

Demonstrate how the calculation of the logit difference of CPR (recoded) = yes versus CPR (recoded) = no is equivalent to the value of the log odds ratio obtained in the Question 2A.

The logit difference of CPR (recoded) =yes versus CPR (recoded) =no is

$\text{logit}(\hat{p}_1) - \text{logit}(\hat{p}_2) = \ln\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) - \ln\left(\frac{\hat{p}_2}{1-\hat{p}_2}\right)$ where $\hat{p}_1$ =P(STA=1|CPR=yes) and $\hat{p}_2$ =P(STA=1 |CPR=no)

$= (\widehat{\propto} + \hat{\beta} cpr * 2) - (\widehat{\propto} + \hat{\beta}\ cpr * 4)$

$= \hat{\beta} cpr * (2 - 4)$

$= -0.5871249 * -2$

$= 1.1742498$

$\approx \mathbf{1.174\ (4\ s.f)}$

Recall that the estimated slope coefficient from the logistic regression from question 2A, which is also the log odds ratio, is 1.17425 ≈ ***1.174*** **(4 s.f)**

Hence, the calculation of the logit difference of CPR (recoded) = yes versus CPR (recoded) = no is equivalent to the value of the log odds ratio obtained in the Question 2A.

Use the results from the logistic regression to obtain the 95% CI for the odds ratio and verify that they are the same limits as obtained in Question 2A

Firstly, recall that the coefficient of the slope of the recoded CPR from question 2B is -0.5871249

The odds ratio from the logistic regression of recoded CPR is

$$OR(CPR = yes, CPR = no) = \frac{Odds(STA = 1|CPR = yes)}{Odds(STA = 1|CPR = no)} = \frac{e^{\widehat{\propto}+\widehat{\beta}*2}}{e^{\widehat{\propto}+\widehat{\beta}*4}} = e^{-2*\widehat{\beta}} =$$

$e^{-2*-0.5871249}$

$= 3.23571$

$\approx 3.236\ (4\ \text{s.f})$

Using R to compute the standard error of the slope coefficient of STA vs recoded CPR

summary(recodedcprlogit)$coefficients[2,2]

```
> summary(recodedcprlogit)$coefficients[2,2]
[1] 0.2021326
```

The standard error of the slope coefficient of STA vs recoded CPR is 0.2021326.

To compute the standard error of the log odds ratio of recoded cpr

$\widehat{SE}(ln[\widehat{OR}(CPR = 2\ vs.\ CPR = 4)])$ = $\left|-2 * \widehat{\mathrm{SE}}(\hat{\beta}_{\mathrm{cpr})}\right|$, where -2 is the change from 4 to 2

= $|-2\ *\ 0.2021326|$

= 0.4042652

95% Confidence interval for the log odds ratio for CPR = 4 (No) vs. CPR = 2 (Yes)

$$ln[\widehat{OR}(CPR = 4\ vs.\ CPR = 2)] \pm 1.96 * \widehat{SE}(ln[\widehat{OR}(CPR = 4\ vs.\ CPR = 2)])$$

95% Confidence interval for the odds ratio for CPR = 4 (No) vs. CPR = 2 (Yes)

$e^{ln[\widehat{OR}(CPR=4\ vs.\ CPR=2)]\pm 1.96*\widehat{SE}(ln[\widehat{OR}(CPR=4\ vs.\ CPR=2)])}$

$= e^{(-2*-0.5871249)\pm 1.96*0.4042652}$

$= (1.46505, 7.14640)$

$\approx (1.465, 7.146)$ (4 s. f)

Recall that the 95% confidence interval for odds ratio in question 2A ≈ (1.465, 7.146) (4 s.f)

Hence, we have verified that using results from the logistic regression to obtain the 95% CI for the odds ratio obtains the same limits as obtained in Question 2A.

Question 2 C

Consider the ICU data and use as the outcome variable vital status (STA) and race (RACE) as covariate.

R code for sta vs race as factor

```
race=data$V5
racenew=as.factor(race)
racemodel=glm(sta~racenew, family=binomial(link = logit))
summary(racemodel)
```

```
Call:
glm(formula = sta ~ racenew, family = binomial(link = logit))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6937  -0.6937  -0.6937  -0.4172   2.2293

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.30174    0.13111  -9.929   <2e-16 ***
racenew2    -0.08456    0.47489  -0.178    0.859
racenew3    -1.09616    0.74979  -1.462    0.144
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 402.97  on 397  degrees of freedom
AIC: 408.97

Number of Fisher Scoring iterations: 4
```

Prepare a table showing the coding of the two dummy variables for RACE using the value RACE=1, white as the reference group.

R code for table showing coding of dummy variables for race

```
contrasts(racenew)
```

```
  2 3
1 0 0
2 1 0
3 0 1
```

Computing the output in a table,

Race (reference group is Race =1)	Race dummied =2	Race dummied=3
Race=1	0	0
Race=2	1	0
Race= 3	1	1

Table 4: Coding of Race with reference to race=1

Show that the estimated log-odds ratios obtained from the cross-tabulation of STA by RACE, using RACE=1 as the reference group, are identical to estimated slope coefficients for the two dummy variables from the logistic regression of STA on RACE.

R code to compute the cross tabulation

```
xtabs(~sta+racenew,data=data)
```

```
    racenew
sta   1   2   3
  0 272  24  22
  1  74   6   2
```

Computing the output in a table,

	Race = 3	Race = 2	Race = 1
STA = 1	2	6	74
STA = 0	22	24	272

Table 4: Cross-tabulation of sta vs race

Using the cross tabulation table 4, for odds ratio for race=2 using race=1 as reference,

$$\widehat{\mathrm{OR}}(\mathrm{CPR}=1 \text{ vs } \mathrm{CPR}=0) = \frac{6*272}{74*24} = \frac{34}{37} = 0.918918 \approx 0.9189 \text{ (4 s.f)}$$

$$ln\left[\widehat{\mathrm{OR}}(\mathrm{CPR}=1 \text{ vs } \mathrm{CPR}=0)\right] = \ln(\frac{34}{37}) = -0.0845573 \approx -0.08456 \text{ (4 s.f)}$$

The log odds ratio from the cross-tabulation for race= 2 using race=1 as reference is -*0.08456* (4 s.f)

For odds ratio for race=3 using race=1 as reference,

$$\widehat{\mathrm{OR}}(\mathrm{CPR}=1 \text{ vs } \mathrm{CPR}=0) = \frac{2*272}{74*22} = \frac{136}{407} = 0.334152 \approx 0.3342 \text{ (4 s.f)}$$

$$ln\left[\widehat{\mathrm{OR}}(\mathrm{CPR}=1 \text{ vs } \mathrm{CPR}=0)\right] = \ln(\frac{136}{407}) = -1.09615 \approx -1.096 \text{ (4 s.f)}$$

The log odds ratio from the cross-tabulation for race= 3 using race=1 as reference is -*1.096* (4 s.f)

Recall using R that the slope coefficients for sta vs race is as follows

```
summary(racemodel)$coefficients[2,1]
[1] -0.08455739
summary(racemodel)$coefficients[3,1]
[1] -1.096158
```

The estimated slope coefficients for the two dummy variables from the logistic regression of STA on race is -0.08455739 ≈ -0.08456 (4 s.f) and -1.096158 ≈ -1.096 (4 s.f)

Hence, that the estimated log-odds ratios obtained from the cross-tabulation of STA by RACE, using RACE=1 as the reference group, are identical to estimated slope coefficients for the two dummy variables from the logistic regression of STA on RACE.

Verify that the estimated standard errors of the estimated slope coefficients for the two dummy variables for RACE are identical to the square root of the sum of the inverse of the cell frequencies from the cross-tabulation of STA by RACE used to calculate the odds ratios.

Square root of the sum of the inverse of the cell frequencies from the cross-tabulation of STA by RACE used to calculate the odds ratios

Recalling table 4 as follows

	Race = 1	Race = 2	Race = 3
STA = 1	74	6	2
STA = 0	272	24	22

Table 4: Cross-tabulation of sta vs race

For race =2 with respect to race=1 as reference group

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{74}+\frac{1}{272}+\frac{1}{6}+\frac{1}{24}} = 0.474892 \approx \underline{\textbf{0.4749 (4 s.f)}}$$

For race =3 with respect to race=1 as reference group

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} = \sqrt{\frac{1}{74}+\frac{1}{272}+\frac{1}{2}+\frac{1}{22}} = 0.75009 \approx \underline{\textbf{0.7501 (4 s.f)}}$$

Recall using R that the slope coefficients for sta vs race is as follows

```
summary(racemodel)$coefficients[2,2]
```

[1] 0.4748928

```
summary(racemodel)$coefficients[3,2]
```

[1] 0.749793

The estimated standard errors of the estimated slope coefficients for the two dummy variables for RACE is $\underline{\approx \mathbf{0.4749}}$ (4 s.f) and $\underline{\approx \mathbf{0.7498}}$ (4 s.f)

Hence, the estimated standard errors of the estimated slope coefficients for the two dummy variables for RACE are identical at 2 significant figures to the square root of the sum of the inverse of the cell frequencies from the cross-tabulation of STA by RACE used to calculate the odds ratios.

Use either set of computations to compute the 95% CI for the odd ratios

95% Confidence interval for the log odds ratio is

$$ln[\widehat{OR}] \pm 1.96 * \widehat{SE}(ln[\widehat{OR}])$$

95% Confidence interval for the odds ratio is

$$e^{ln[\widehat{OR}] \pm 1.96 * \widehat{SE}(ln[\widehat{OR}])}$$

Using the cross tabulation results of STA vs RACE earlier,

For race =2 with respect to race=1 as reference group

The 95% confidence interval for log odds ratio is $-0.0845573 \pm 1.96\ (0.474892)$

= (-1.01534, 0.846231)

≈ (-1.015, 0.8462) (4 s.f)

The 95% confidence interval for odds ratio is = $e^{-0.0845573 \pm 1.96\,(0.474892)}$

= (0.362277, 2.33084)

≈ (0.3623, 2.331) (4 s.f)

For race =3 with respect to race=1 as reference group

The 95% confidence interval for log odds ratio is $-1.09615 \pm 1.96\,(0.75009)$

= (-2.56632, 0.374026)

≈ (-2.566, 0.3740) (4 s.f)

The 95% confidence interval for odds ratio is = $e^{-1.09615 \pm 1.96\,(0.75009)}$

= (0.0768172,1.45357)

≈ (0.07682, 1.454) (4 s.f)

Using R for a quick check (extra)

```
exp(summary(racemodel)$coefficients[2,1]-1.96*summary(racemodel)$coefficients[2,2])
[1] 0.3622766
 exp(summary(racemodel)$coefficients[2,1]+1.96*summary(racemodel)$coefficients[2,2])
[1] 2.330849
 exp(summary(racemodel)$coefficients[3,1]-1.96*summary(racemodel)$coefficients[3,2])
[1] 0.07686134
 exp(summary(racemodel)$coefficients[3,1]+1.96*summary(racemodel)$coefficients[3,2])
[1] 1.452718
```

Question 2 D

Prepare a table showing the coding of three dummy variables based on the empirical quartiles of AGE using the first quartiles as the reference group.

Using R to obtain the empirical quartiles of age

```
quantile(age)
 0%  25%  50%  75% 100%
 16   47   61   71   92
```

Computing the above result into a table

Quartiles	Q1	Q2	Q3
Age	47	61	71

Table 5: Empirical quartiles of age

The coding of three dummy variables based on the empirical quartiles of AGE using the first quartiles as the reference group

Age groups based on empirical quartiles obtained	Age dummy 1	Age dummy 2	Age dummy 3
Q1 = (15,47]	0	0	0
Q2 = (47,61]	1	0	0
Q3 = (61,71]	0	1	0
Q4 = (71,92]	0	0	1

Fit the logistic regression of STA on AGE as recoded into these design variables

```
data$agenew<-cut(age,breaks=c(15,47,61,71,92), ordered = TRUE)
agenew=data$agenew
table(data$agenew)
```

```
(15,47] (47,61] (61,71] (71,92]
   108      93     100      99
```

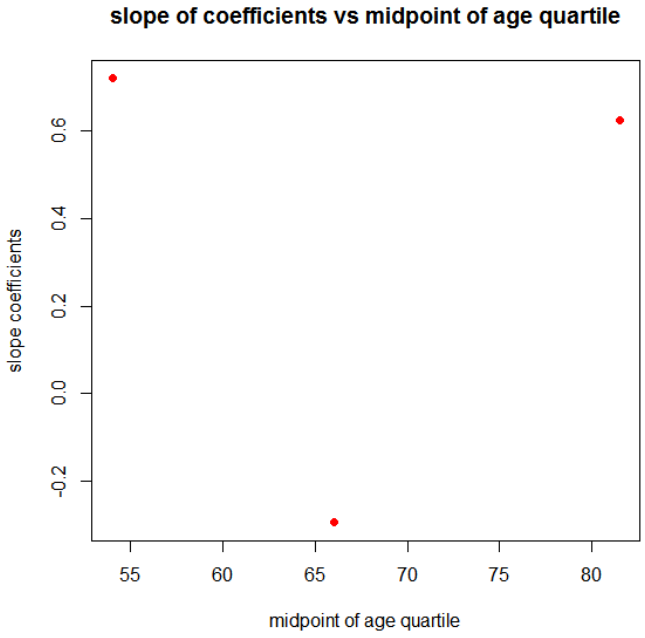
```
agemodel <- glm(sta ~ agenew, data = data, family = binomial(link = logit))
```

```
> summary(agemodel)

Call:
glm(formula = sta ~ agenew, family = binomial(link = logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8154  -0.7913  -0.6300  -0.4635   2.1374

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4061     0.1307 -10.755  < 2e-16 ***
agenew.L      0.7208     0.2722   2.648  0.00809 **
agenew.Q     -0.2952     0.2615  -1.129  0.25896
agenew.C      0.6246     0.2503   2.495  0.01258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 391.57  on 396  degrees of freedom
AIC: 399.57

Number of Fisher Scoring iterations: 4
```

Plot the three estimated slope coefficients versus the mid-point of the respective age quartile

Using R

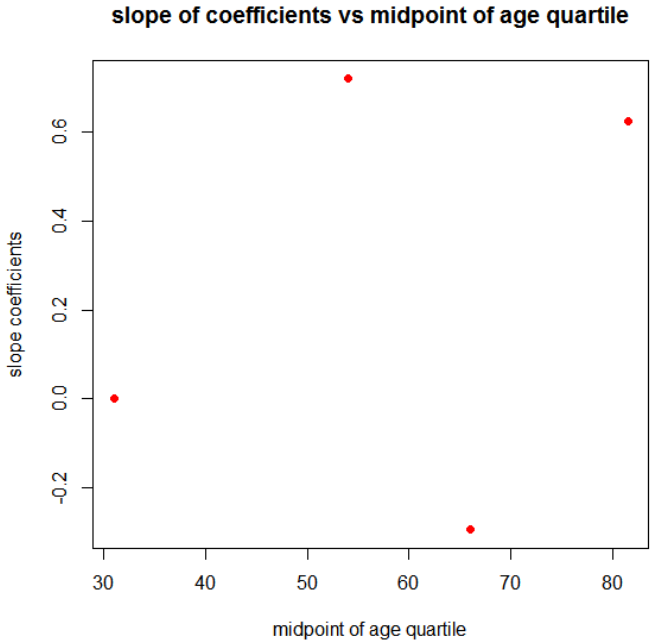
```
plot(midpoint2,summary(agemodel)$coefficients[2:4,1],ylab="slope coefficients",xlab="midpoint of age quartile",pch=16,col="red",main="slope of coefficients vs midpoint of age quartile")
```



Plot a fourth point a value of zero at the mid-point of the first quartile of age

R code

```
fourthpoint=c(0,summary(agemodel)$coefficients[2:4,1])
midpoint3=c(median(15:47),median(47:61),median(61:71),median(71:92))
plot(midpoint3,fourthpoint,ylab="slope coefficients",xlab="midpoint of age
quartile",pch=16,col="red",main="slope of coefficients vs midpoint of age quartile")
```



Dose this plot suggest that the logit is linear in age?

No

Question 2 E

Consider the logistic regression of STA on CRN and AGE

The logistic regression of STA on CRN and AGE is

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 crn$$

For patient i, CRN =0 if patient has no history of chronic renal failure prior to ICU admission and CRN =1 if patient has history of chronic renal failure prior to ICU admission. AGE variable is the age of the patient, which is a continuous variable. $\hat{p}_i$ is the estimated probability of death of patient i upon admission to ICU. $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ are parameter estimates for the intercept and slope coefficients respectively.

To check if age is a confounder for the association of CRN and STA, we will compare the logistic equation of STA vs CRN and AGE (which is the main effect model) and logistic regression of STA vs CRN only.

Using R,

```
crn=data$V8
crnagemodel <- glm(sta ~ age+crn, data = data, family = binomial(link = logit))
> summary(crnagemodel)

Call:
glm(formula = sta ~ age + crn, family = binomial(link = logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9646  -0.7261  -0.6428  -0.4597   2.1366

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.545949   0.452103  -5.631 1.79e-08 ***
age          0.019529   0.007099   2.751  0.00594 **
crn          0.303831   0.399888   0.760  0.44738
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 396.40  on 397  degrees of freedom
AIC: 402.4

Number of Fisher Scoring iterations: 4
```

Consider CRN to be the risk factor and show that AGE is a confounder of the association of CRN with STA.

R code to model STA vs CRN (History of Chronic Renal Failure, O = No, 1 = Yes) only:

```
crnmodel <- glm(sta ~ crn, data = data, family = binomial(link = logit))
```

```
> summary(crnmodel)

Call:
glm(formula = sta ~ crn, family = binomial(link = logit), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8067  -0.6639  -0.6639  -0.6639   1.8003

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4001     0.1316 -10.641   <2e-16 ***
crn           0.4446     0.3947   1.126     0.26
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 404.60  on 398  degrees of freedom
AIC: 408.6

Number of Fisher Scoring iterations: 4
```

Percentage change in coefficient of CRN after AGE is added to the model is

$\frac{0.303831-0.4446}{0.4446}$ x 100 = -31.6619%

Hence, age is a confounder of the association of CRN with STA because after age is included to the model, the estimated coefficient risk factor CRN decreased by approximately 31.66% (4 s.f). Also, the association between AGE and STA is statistically significant (p value =0.00594).

Addition of the interaction of AGE by CRN presents an interesting modelling dilemma.

For interaction model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 crn + \hat{\beta}_3 age * crn$$

R code to model STA vs CRN, AGE and interaction term CRN * AGE

crnageinteractionmodel = glm(sta ~ age+crn+age*crn, data = data, family = binomial(link = logit))

```
> summary(crnageinteractionmodel)

Call:
glm(formula = sta ~ age + crn + age * crn, family = binomial(link = logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0786  -0.7270  -0.6369  -0.4295   2.1946

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.747539   0.482426  -5.695 1.23e-08 ***
age          0.022822   0.007532   3.030  0.00244 **
crn          3.184494   1.767499   1.802  0.07159 .
age:crn     -0.044561   0.027315  -1.631  0.10281
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 393.72  on 396  degrees of freedom
AIC: 401.72

Number of Fisher Scoring iterations: 4
```

Examine the main effects only and interaction models graphically.

For main effect model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 crn$$

For CRN = 0, $\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 age$

$$\hat{p}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 age)}}$$

For CRN = 1, $\text{logit}(\hat{p}_i) = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1) age$

$$\hat{p}_i = \frac{1}{1+e^{-((\hat{\beta}_0+\hat{\beta}_2)+(\hat{\beta}_1) age)}}$$

For interaction model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 crn + \hat{\beta}_3 age * crn$$

For CRN = 0, $\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 age$

$$\hat{p}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 age)}}$$

For CRN = 1, $\text{logit}(\hat{p}_i) = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) age$

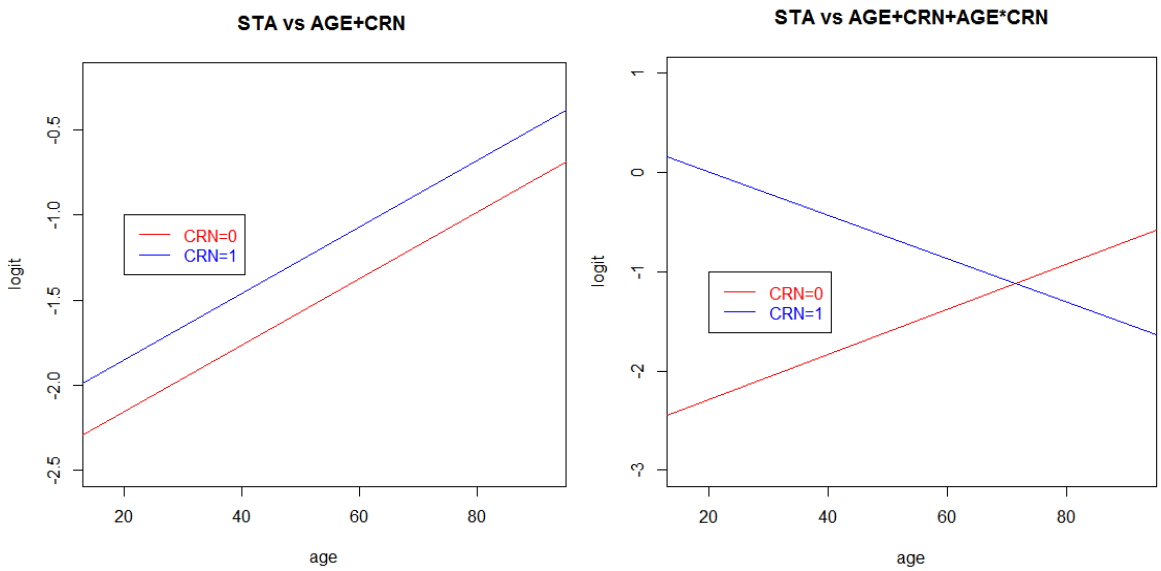
$$\hat{p}_i = \frac{1}{1+e^{-((\hat{\beta}_0+\hat{\beta}_2)+(\hat{\beta}_1+\hat{\beta}_3) age)}}$$

R code for logit plot for main effect model of STA vs AGE and CRN

```
plot(age,sta,type="n",ylab="logit",main="STA vs AGE+CRN ",ylim=c(-2.5,-0.2))
abline(crnagemodel$coefficients[1], crnagemodel $coefficients[2],col="red")
abline(crnagemodel$coefficients[1]+ crnagemodel$coefficients[3], crnagemodel$coefficients[2],
col="blue")
legend(20,-1,c("CRN=0","CRN=1"),text.col=c("red","blue"),col=c("red","blue"),lwd=1)
```

R code for logit plot for interaction model of STA VS AGE,CRN and CRN*AGE

```
plot(age,sta,type="n",ylab="logit",main="STA vs AGE+CRN+AGE*CRN",ylim=c(-3,1))
abline(crnageinteractionmodel$coefficients[1],crnageinteractionmodel$coefficients[2],col="red")
abline(crnageinteractionmodel$coefficients[1]+crnageinteractionmodel$coefficients[3],crnageintera
ctionmodel$coefficients[2]+crnageinteractionmodel$coefficients[4],col="blue")
legend(20,-1,c("CRN=0","CRN=1"),text.col=c("red","blue"),col=c("red","blue"),lwd=1)
```



Using the graphical results and any significance tests you feel are needed, select the best model (main effects or interaction) and justify your choice.

Using R computing Anova chi sq test to select best model

```
anova(crnmodel,crnagemodel,crnageinteractionmodel,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: sta ~ crn
Model 2: sta ~ age + crn
Model 3: sta ~ age + crn + age * crn
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       398     404.60
2       397     396.40  1   8.2047 0.004178 **
3       396     393.72  1   2.6840 0.101363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion is to use model 2 which is STA vs AGE+CRN (p-value = 0.004178). Does not seems justified to add interaction and adding age seems to useful. Also, using results earlier, AIC for model 2 (AIC: 402.4) is approximately same as model 3 (AIC: 401.72). Hence, choose the simpler model which is model 2. In additional, from the earlier plot for STA vs AGE+CRN without interaction term also suggests that the model has no interaction term for CRN=1 and CRN=0. Thus, model 2 is best.

Estimate relevant odds ratios

Recall earlier the results for STA vs AGE+CRN

```
crnagemodel$coefficients[3]
```

```
      crn
0.3038315
```

Estimated odds ratio for CRN = 1 vs. CRN = 0 adjusted for AGE:

$$\widehat{\mathrm{OR}}(\mathrm{CRN} = 1 \text{ vs. } \mathrm{CRN} = 0) = e^{0.3038315} = 1.35504 \approx 1.355 \text{ (4 s.f)}$$

Repeat this analysis of confounding and interaction for a model that includes CPR as the risk factor and AGE as the potential confounding variable.

The logistic regression of STA on CPR and AGE is

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 cpr$$

For patient i, CPR =0 if patient did not receive CPR prior to ICU admission and CPR =1 if patient receives CPR prior to ICU admission. AGE variable is the age of the patient, which is a continuous variable. $\hat{p}_i$ is the estimated probability of death of patient i upon admission to ICU. $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ are parameter estimates for the intercept and slope coefficients respectively.

To check if age is a confounder for the association of CPR and STA, we will compare the logistic equation of STA vs CPR and AGE (which is the main effect model) and logistic regression of STA vs CPR only.

Using R, the logistic regression of STA on CPR and AGE is

```
cpragemodel <- glm(sta ~ age+cpr, data = data, family = binomial(link = logit))
```

```
> summary(cpragemodel)

Call:
glm(formula = sta ~ age + cpr, family = binomial(link = logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1443  -0.7064  -0.6175  -0.4293   2.1959

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.716324   0.466998  -5.817 6.01e-09 ***
age          0.021017   0.007211   2.914  0.00356 **
cpr          1.229816   0.412101   2.984  0.00284 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 388.69  on 397  degrees of freedom
AIC: 394.69

Number of Fisher Scoring iterations: 4
```

R code to model STA vs CPR only:

```
cprmodel<- glm(sta ~ cpr, data = data, family = binomial(link = logit))
```

```
> summary(cprmodel)

Call:
glm(formula = sta ~ cpr, family = binomial(link = logit), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0579  -0.6457  -0.6457  -0.6457   1.8278

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4619     0.1327 -11.021  < 2e-16 ***
cpr           1.1742     0.4043   2.905  0.00368 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 398.01  on 398  degrees of freedom
AIC: 402.01

Number of Fisher Scoring iterations: 4
```

Percentage change in coefficient of CPR after age is added to the model is

$\frac{1.229816-1.1742}{1.1742}$x 100 = 4.73650%

Hence, age may not be a confounder of the association of CPR with STA because after age is included to the model, the estimated coefficient risk factor CPR increased by approximately 4.737% (4 s.f) only.

Now adding the interaction CPR* AGE into the main effect model

For interaction model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 cpr + \hat{\beta}_3 age * cpr$$

R code to model STA vs CPR, AGE and interaction term CPR * AGE

```
cprageinteractionmodel = glm(sta ~ age+cpr+age*cpr, data = data, family = binomial(link = logit))
```

```
> summary(cprageinteractionmodel)

Call:
glm(formula = sta ~ age + cpr + age * cpr, family = binomial(link = logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5096  -0.6894  -0.6292  -0.4874   2.0802

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.317364   0.460639  -5.031 4.89e-07 ***
age          0.014513   0.007272   1.996   0.0460 *
cpr         -5.642967   3.178742  -1.775   0.0759 .
age:cpr      0.115547   0.051580   2.240   0.0251 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 378.56  on 396  degrees of freedom
AIC: 386.56

Number of Fisher Scoring iterations: 5
```

Examining the main effect and interaction model graphically,

For main effect model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 cpr$$

For CPR= 0, $\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 age$

$$\hat{p}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 age)}}$$

For CPR = 1, $\text{logit}(\hat{p}_i) = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1) age$

$$\hat{p}_i = \frac{1}{1+e^{-((\hat{\beta}_0+\hat{\beta}_2)+(\hat{\beta}_1) age)}}$$

For interaction model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 age + \hat{\beta}_2 cpr + \hat{\beta}_3 age * cpr$$

For CPR = 0, $\text{logit}(\hat{p}_i) = \hat{\beta}_0 + \hat{\beta}_1 age$

$$\hat{p}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 age)}}$$

For CPR = 1, $\text{logit}(\hat{p}_i) = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) age$

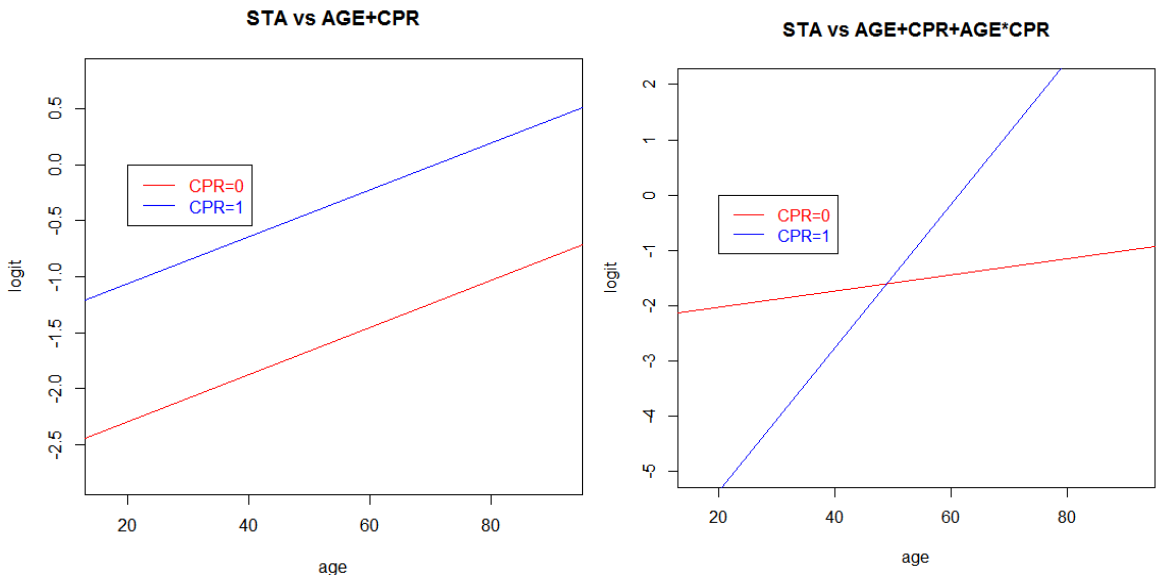
$$\hat{p}_i = \frac{1}{1+e^{-((\hat{\beta}_0+\hat{\beta}_2)+(\hat{\beta}_1+\hat{\beta}_3) age)}}$$

R code for logit plot for main effect model of STA vs AGE and CPR

```
plot(age,sta,type="n",ylab="logit",main="STA vs AGE+CPR",ylim=c(-2.8,0.8))
abline(cpragemodel$coefficients[1], cpragemodel $coefficients[2],col="red")
abline(cpragemodel$coefficients[1]+ cpragemodel$coefficients[3], cpragemodel$coefficients[2],
col="blue")
legend(20,0,c("CPR=0","CPR=1"),text.col=c("red","blue"),col=c("red","blue"),lwd=1)
```

R code for logit plot for interaction model of STA VS AGE,CPR and CPR*AGE

```
plot(age,sta,type="n",ylab="logit",main="STA vs AGE+CPR+AGE*CPR",ylim=c(-5,2))
abline(cprageinteractionmodel$coefficients[1],cprageinteractionmodel$coefficients[2],col="red")
abline(cprageinteractionmodel$coefficients[1]+cprageinteractionmodel$coefficients[3],cprageintera
ctionmodel$coefficients[2]+cprageinteractionmodel$coefficients[4],col="blue")
legend(20,0,c("CPR=0","CPR=1"),text.col=c("red","blue"),col=c("red","blue"),lwd=1)
```



Using R to compute likelihood ratio test using Anova chi sq test to select best model

```
anova(cprmodel,cpragemodel,cprageinteractionmodel,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: sta ~ cpr
Model 2: sta ~ age + cpr
Model 3: sta ~ age + cpr + age * cpr
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       398     398.01
2       397     388.69  1   9.3192 0.002268 **
3       396     378.56  1  10.1369 0.001453 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion: Both model 2 (p value = 0.001453) and 3 (p value=0.002268) are statistically significant. Checking the AIC, AIC for model 2 (AIC: 394.69) is approximately same as model 3 (AIC: 386.56). Hence, choose the simpler model which is model 2. In additional, from the earlier plot for STA vs age+cpr without interaction term also doesn't seems justified to add the interaction term.
We also deduced earlier that age might not be a confounder for STA vs CPR. AIC for model (AIC: 402.01) is also not much difference from model 2 and 3.
Thus, model 1 (STA vs CPR only) in this case seems like the best option.

Estimating the odds ratio for model 1,

Recall earlier the results for STA vs CPR

```
cprmodel$coefficients[2]
```

```
     cpr
1.17425
```

The estimated odds ratio for CPR = 1 vs. CPR = 0:

$$\widehat{\mathrm{OR}}(\mathrm{CPR} = 1 \text{ vs. } \mathrm{CPR} = 0) = e^{1.17425} = 3.23571 \approx 3.236 \text{ (4 s.f)}$$

Question 2 F

Consider an analysis for confounding and interaction for the model with STA as the outcome CAN as the risk factor, and TYP as the potential confounding variable. Perform this analysis using logistic regression modelling.

The logistic regression of STA on CAN and TYP is

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 typ + \hat{\beta}_2 can$$

For patient i, CAN =0 if patient has no cancer as part of present problem and CAN =1 if patient has cancer as part of present problem. TYP variable is the type of admission of patient with 1 as emergency and 0 as elective. It is a dichotomous variable. $\hat{p}_i$ is the estimated probability of death of patient i upon admission to ICU. $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ are parameter estimates for the intercept and slope coefficients respectively.

To check if TYP is a confounder for the association of CAN and STA, we will compare the logistic equation of STA vs CAN and TYP (which is the main effect model) and logistic regression of STA vs CAN only.

Using R, the logistic regression of STA on CAN and TYP is

```
can=data$V7
data$can=can
typ=data$V14
data$typ=typ
cantypmodel <- glm(sta ~ typ+can, data = data, family = binomial(link = logit))
```

```
> summary(cantypmodel)

Call:
glm(formula = sta ~ typ + can, family = binomial(link = logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6443  -0.7639  -0.7639  -0.1820   2.8667

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.0925     0.6258  -6.540 6.15e-11 ***
typ           3.0101     0.6269   4.801 1.58e-06 ***
can           2.1348     0.5962   3.581 0.000343 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 358.66  on 397  degrees of freedom
AIC: 364.66

Number of Fisher Scoring iterations: 6
```

R code to model STA vs CAN only:

```
canmodel<- glm(sta ~ can, data = data, family = binomial(link = logit))
> summary(canmodel)

Call:
glm(formula = sta ~ can, family = binomial(link = logit), data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8141  -0.6618  -0.6618  -0.6618   1.8034

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4072     0.1324 -10.627   <2e-16 ***
can           0.4729     0.3797   1.246    0.213
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 404.34  on 398  degrees of freedom
AIC: 408.34

Number of Fisher Scoring iterations: 4
```

Percentage change in coefficient of CAN after TYP is added to the model is

$\frac{2.1348-0.4729}{0.4729}$ X 100 = 351.427%

Hence, the estimated coefficient risk factor CAN increased greatly by approximately 351.4% (4 s.f) after TYP is added to model controlling TYP, which indicates that TYP is a confounder.

Now adding the interaction term CAN*TYP into the main effect model

For interaction model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 typ + \hat{\beta}_2 can + \hat{\beta}_3 typ * can$$

R code to model STA vs CAN, TYP and interaction term CAN*TYP

```
cantypinteractionmodel = glm(sta ~ typ+can+typ*can, data = data, family = binomial(link = logit))
```

```
> summary(cantypinteractionmodel)

Call:
glm(formula = sta ~ typ + can + typ * can, family = binomial(link = logit),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7941  -0.7601  -0.7601  -0.2169   2.7427

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.7377     0.7155  -5.224 1.75e-07 ***
typ           2.6439     0.7289   3.627 0.000286 ***
can           1.5782     0.9400   1.679 0.093182 .
typ:can       0.9019     1.2361   0.730 0.465637
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 405.81  on 399  degrees of freedom
Residual deviance: 358.13  on 396  degrees of freedom
AIC: 366.13

Number of Fisher Scoring iterations: 6
```

Examining the main effect and interaction model graphically,

For main effect model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 typ \ + \hat{\beta}_2 can$$

For CAN= 0, $\text{logit}(\hat{p}_i) = \ \hat{\beta}_0 + \hat{\beta}_1 typ$

$$\hat{p}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 typ)}}$$

For CAN = 1, $\text{logit}(\hat{p}_i) = \ (\hat{\beta}_0 + \hat{\beta}_2) + \ (\hat{\beta}_1) typ$

$$\hat{p}_i = \frac{1}{1+e^{-((\hat{\beta}_0+\hat{\beta}_2)+ (\hat{\beta}_1) typ)}}$$

For interaction model

$$\text{logit}(\hat{p}_i) = \ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = \hat{\beta}_0 + \hat{\beta}_1 typ + \hat{\beta}_2 can \ + \hat{\beta}_3 typ * can$$

For CAN = 0, $\text{logit}(\hat{p}_i) = \ \hat{\beta}_0 + \hat{\beta}_1 typ$

$$\hat{p}_i = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 age)}}$$

For CAN = 1, $\text{logit}(\hat{p}_i) = \ (\hat{\beta}_0 + \hat{\beta}_2) + \ (\hat{\beta}_1 + \hat{\beta}_3)\ typ$

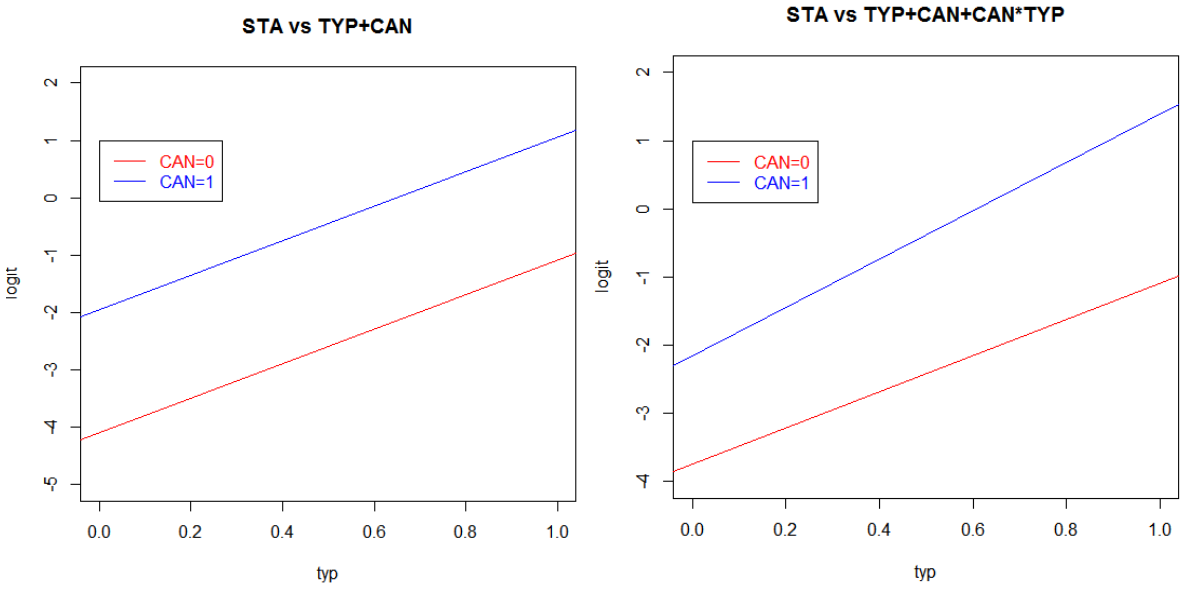
$$\hat{p}_i = \frac{1}{1+e^{-((\hat{\beta}_0+\hat{\beta}_2)+ (\hat{\beta}_1+\hat{\beta}_3) typ)}}$$

R code for logit plot for main effect model of STA vs TYP and CAN

```
plot(typ,sta,type="n",ylab="logit",main="STA vs TYP+CAN",ylim=c(-5,2))
abline(cantypmodel$coefficients[1], cantypmodel $coefficients[2],col="red")
abline(cantypmodel$coefficients[1]+ cantypmodel$coefficients[3], cantypmodel$coefficients[2],
col="blue")
legend(0,1,c("CAN=0","CAN=1"),text.col=c("red","blue"),col=c("red","blue"),lwd=1)
```

R code for logit plot for interaction model of STA VS AGE,CPR and CPR*AGE

```
plot(typ,sta,type="n",ylab="logit",main="STA vs TYP+CAN+CAN*TYP",ylim=c(-4,2))
abline(cantypinteractionmodel$coefficients[1], cantypinteractionmodel $coefficients[2],col="red")
abline(cantypinteractionmodel$coefficients[1]+ cantypinteractionmodel$coefficients[3],
cantypinteractionmodel$coefficients[2]+ cantypinteractionmodel$coefficients[4], col="blue")
legend(0,1,c("CAN=0","CAN=1"),text.col=c("red","blue"),col=c("red","blue"),lwd=1)
```



Using R to compute likelihood ratio test using Anova chi sq test to select best model

```
anova(canmodel,cantypmodel,cantypinteractionmodel,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: sta ~ can
Model 2: sta ~ typ + can
Model 3: sta ~ typ + can + typ * can
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       398     404.34
2       397     358.66  1   45.677 1.394e-11 ***
3       396     358.13  1    0.533    0.4654
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusion is to use model 2 which is STA vs CAN+TYP (p value ≈ 0). Does not seems justified to add interaction and adding TYP seems to useful. Also, using results earlier, AIC for model 2 (AIC: 364.66) is approximately same as model 3 (AIC: 366.13). Hence, choose the simpler model which is model 2.

In additional, from the earlier plot for STA vs CAN+TYP without interaction term also suggests that the model has no interaction term for CAN=1 and CAN=0. Thus, model 2 is best.

Estimating the odds ratio for model 2,

Recall earlier the results for STA vs TYP+CAN

```
cantypmodel$coefficients[3]
```

```
      can
2.134818
```

The estimated odds ratio for CAN = 1 vs. CAN = 0 adjusted for TYP:

$$\widehat{\mathrm{OR}}(\mathrm{CAN} = 1 \text{ vs. } \mathrm{CAN} = 0) = e^{2.134818} = 8.45550 \approx 8.456 \text{ (4 s.f)}$$

Section II

Question 1

Describe what are the differences between odds ratio and relative risk, in terms of study design, interpretations etc.

Odds Ratio (OR) and Relative Risk (RR) both assess the strength of association between the exposed, or non-exposed, and outcome of interest which gives indicatives of how more or less likely a group is likely to develop disease compared to another group.

Relative risk can be thought of as a constant relative risk across strata or multiplicative effect, leading to a multiplicative model. In addition, Relative risk is more useful if reason behind why exposure causes the disease is not independent for non-exposed cases. Otherwise, biologically, the interpretation would not make much sense.

Relative risk is used with cohort studies only but not case control as population at risk is unknown. Relative risk depends on the sampling fraction used whereas odds ratio is independent. Hence, odds ratio can be used with both case control and cohort studies.

The interpretations are similar in terms of quantified association. Example, both OR and RR =1 indicates no association, <1 indicates negative association and vice versa. However the theoretical interpretation is slightly different. Example RR =5 represents people are 5 times more likely to have the outcome when compared to people who were not exposed. But OR = 5 represents people who have the outcome is 5 times more likely to be exposed than not exposed.

OR is also always greater than or approximately equal to RR because $OR = \frac{1-P_0}{1-P_1}\ RR$

Question 2

Describe three methods that commonly used in epidemiological data analysis to control confounding effects.

Confounding factors, if not controlled for, can cause bias in the estimate of the impact of the exposure being studied. There are many ways to control confounding effects. At the design stage, methods such as matching can be used. But during the epidemiological data analysis phase, the three most commonly used methods to control confounding effects are regression, stratification and standardization.

Stratification allows the association between exposure and outcome to be examined within different strata of the confounding variable. Stratification only works best if there are not a lot of strata and if only 1 or 2 confounders have to be controlled. It is commonly used to adjust for confounder age variable. Initially the strength of the association is measured within each stratum of the confounding variable. Assuming each stratum rates are approximately uniform, they can be pooled to provide a summary estimate of the relative risk controlling confounder. Within each stratum, the confounder cannot confound because it does not vary, unless of cause inadequacy in adjustments made. But the disadvantage of stratification is that the more stratified the sample is, the smaller each stratum becomes, and hence the discrimination power to detect associations is reduced.

If the number of potential confounders or the level of their grouping is large, control using regression method is better as it can control many confounders at the same time assuming sample size is large enough. However in general, stratification requires less assumption than control via regression. Example controlling via regression requires between the confounders and the outcome. Hence, it may be prone to wrong assumption of relationship.

There are many methods to control using regression methods such as logistic, poisson, cox etc. Control of confounder is done by include confounders in the model adjusting for them. Selecting of model is based on characteristic of data sample, variables and analysis objective.

Standardisation is an example of stratification. There are two ways to do so: direct standardization and indirect method. Standardization gives a quantitative measure of the difference in rates between the study cohort and a standard/reference population or comparison group that is free

from effects of potential confounding variables. Reason to standardize between two groups is because crude rates may be misleading if they are unadjusted for confounders. However, standardization does not work if there is interaction effect.

However, it is crucial to note residual confounding can still exist even after adjustments are made. Confounding variables can be misclassified, categories of model being too broad, variables missing in the model from inadequacy in adjustments etc. Hence, selection of method and careful adjustment using any methods is crucial.

Question 3

Describe the two criteria to assess the adequacy of a logistic regression model

Two criteria to assess the adequacy of a logistic regression model is its discrimination and calibration (goodness-of-fit)

Discrimination measures the model's ability to discriminate between those subjects who experience the outcome of interest compared to those who do not. The best model is in which true positive (sensitivity) and true negative (specificity) do not overlap on the plot of sensitivity vs false positive. The AUROC curve can be used to measure the discrimination power by looking at the area under it. AUROC =0.5 for example suggest no discrimination which meant the model selected is not good. The higher the discrimination ability, the better the model is.

Calibration is a measure to describe how closely the predicted probabilities are fitted, or basically how well the model describes the response variables. A value exceeding 20 is considered an index of poor calibration. Hosmer-Lemeshow test is commonly used to measure calibration as it allows for any number of explanatory variables, and they can be categorical or continuous. Hence, it is also used in logistic regression. For matched case-control design, McNemar's Test can be used.