Exercise 6 - solution

Paul Blanche

Exercise A

Part I

Question 1

We first load the data and look at the “summary”, as always.

```r
load(url("http://paulblanche.com/files/MI.rda"))
summary(MI)
```

```
##        id            mi               oc            tobacco     
##  Min.   :  1   Min.   :0.0000   Min.   :0.0000   Min.   :1.000  
##  1st Qu.:113   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000  
##  Median :225   Median :0.0000   Median :0.0000   Median :2.000  
##  Mean   :225   Mean   :0.3318   Mean   :0.4454   Mean   :1.742  
##  3rd Qu.:337   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:2.000  
##  Max.   :449   Max.   :1.0000   Max.   :1.0000   Max.   :3.000  
##                                                                 
##       age             weight           height           bmi       
##  Min.   : 15.00   Min.   : 33.00   Min.   :138.0   Min.   :11.36  
##  1st Qu.: 33.00   1st Qu.: 51.00   1st Qu.:160.0   1st Qu.:18.67  
##  Median : 44.00   Median : 64.00   Median :166.0   Median :23.18  
##  Mean   : 45.62   Mean   : 66.07   Mean   :165.2   Mean   :24.38  
##  3rd Qu.: 56.00   3rd Qu.: 79.00   3rd Qu.:171.0   3rd Qu.:29.17  
##  Max.   :100.00   Max.   :128.00   Max.   :184.0   Max.   :47.78  
##                   NA's   :12                       NA's   :12     
##     history        hypertension   
##  Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000  
##  Mean   :0.1199   Mean   :0.3541  
##  3rd Qu.:0.0000   3rd Qu.:1.0000  
##  Max.   :1.0000   Max.   :1.0000  
##  NA's   :7
```

We can see that 3 variables have missing values: weight (n=12), bmi (n=12) and history (n=7). It does not matter for this exercise because we will not use these variables.

Question 2

We first create and add to the data a factor variable named **Smoke**, which explicitly indicates the smoking status of each woman. We then use summary again to check that the variable has been added to the dataset and see how many observations we have in each group.

```
MI$Smoke <- factor(MI$tobacco,
                   levels=c(1,2,3),
                   labels=c("never smoked",
                            "current smoker",
                            "former smoker"))
summary(MI)
```

```
##        id             mi               oc            tobacco     
##  Min.   :  1   Min.   :0.0000   Min.   :0.0000   Min.   :1.000  
##  1st Qu.:113   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:1.000  
##  Median :225   Median :0.0000   Median :0.0000   Median :2.000  
##  Mean   :225   Mean   :0.3318   Mean   :0.4454   Mean   :1.742  
##  3rd Qu.:337   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:2.000  
##  Max.   :449   Max.   :1.0000   Max.   :1.0000   Max.   :3.000  
##                                                                 
##       age             weight           height           bmi       
##  Min.   : 15.00   Min.   : 33.00   Min.   :138.0   Min.   :11.36  
##  1st Qu.: 33.00   1st Qu.: 51.00   1st Qu.:160.0   1st Qu.:18.67  
##  Median : 44.00   Median : 64.00   Median :166.0   Median :23.18  
##  Mean   : 45.62   Mean   : 66.07   Mean   :165.2   Mean   :24.38  
##  3rd Qu.: 56.00   3rd Qu.: 79.00   3rd Qu.:171.0   3rd Qu.:29.17  
##  Max.   :100.00   Max.   :128.00   Max.   :184.0   Max.   :47.78  
##                   NA's   :12                       NA's   :12     
##     history        hypertension               Smoke    
##  Min.   :0.0000   Min.   :0.0000   never smoked  :215  
##  1st Qu.:0.0000   1st Qu.:0.0000   current smoker:135  
##  Median :0.0000   Median :0.0000   former smoker : 99  
##  Mean   :0.1199   Mean   :0.3541                       
##  3rd Qu.:0.0000   3rd Qu.:1.0000                       
##  Max.   :1.0000   Max.   :1.0000                       
##  NA's   :7                                             
```

We can see that 215 have never smoked, 135 are current smokers and 99 former smokers. We now estimate a “simple” (univariate) logistic model to investigate the association between myocardial infarction (**mi**) and smoking status (**Smoke**).

```
fit1 <- glm(mi~Smoke, data=MI, family=binomial)
summary(fit1)
```

```
##
## Call:
## glm(formula = mi ~ Smoke, family = binomial, data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2735  -1.0842  -0.5868   1.0842   1.9206
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -1.6721     0.1869  -8.946  < 2e-16 ***
## Smokecurrent smoker    1.4490     0.2548   5.686  1.3e-08 ***
## Smokeformer smoker     1.8953     0.2754   6.882  5.9e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 570.66  on 448  degrees of freedom
## Residual deviance: 509.22  on 446  degrees of freedom
## AIC: 515.22
##
## Number of Fisher Scoring iterations: 3
```

The above summary provides us with the parameter estimates. They have an interpretation as either a "log odds" (Intercept) or as the log of an odds ratio (the others parameters). To facilitate the interpretation, we asked for the "formatted" results using the **publish()** function of the **Publish** package.

```
library(Publish)
```

```
## Loading required package: prodlim
```

```
publish(fit1)
```

```
##  Variable          Units OddsRatio        CI.95 p-value
##     Smoke   never smoked       Ref
##           current smoker      4.26  [2.58;7.02]  <1e-04
##            former smoker      6.65 [3.88;11.42]  <1e-04
```

We can now interpret the resuls easily:

- The risk of myocardial infarction (MI) is (significantly) higher for current smokers than for women who have never smoked (OR=4.26, 95% CI=[2.58;7.02], p-value<0.0001).

- The risk of myocardial infarction (MI) is (significantly) higher for former smokers than for women who have never smoked (OR=6.65, 95% CI=[3.88;11.42], p-value<0.0001).

Note however, that (surprisingly) the risk is estimated higher for former smokers than for current smokers, as the odds ratio is higher when comparing former smokers to women who have never smoked than when comparing current smokers to women who have never smoked. The odds ratio to compare former smoker to current smoker can actually be computed as 6.65/4.26=1.56.

Note that in the output of the summary function, we could see the estimated log odds ratios 1.4490 and 1.8953, from which we can deduce the odds ratio exp(1.4490)=4.26 and exp(1.8953)=6.65.

The above results are actually just another way of looking at the following frequency table.

```
table(MI$Smoke,MI$mi)
```

```
##
##                      0   1
##   never smoked     181  34
##   current smoker    75  60
##   former smoker     44  55
```

For instance the odds ratio (and log odds ratio) to compare the risk of MI between women who are current smokers and women who have never smoked can be found as follows:

```
60*181/(75*34)
```

```
## [1] 4.258824
```

```
log(60*181/(75*34))
```

```
## [1] 1.448993
```

And we regognize the results already seen above (up to the “default” rounding when printing the results).

We now make all-pairwise comparisons between the smoking groups and adjust for multiple testing. Here we make three comparisons and we adjust for multiple testing because we want to make sure that if we conclude to any association, then the risk of this conclusion to be incorrect is not “too large”, i.e. not larger than 5%.

```
library(multcomp)
```

```
## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS
```

```
##
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':
##
##     geyser
```

```
Res1 <- glht(fit1, mcp(Smoke="Tukey"))
summary(Res1)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: glm(formula = mi ~ Smoke, family = binomial, data = MI)
##
## Linear Hypotheses:
##                                    Estimate Std. Error z value Pr(>|z|)
## current smoker - never smoked == 0   1.4490     0.2548   5.686   <1e-04 ***
## former smoker - never smoked == 0    1.8953     0.2754   6.882   <1e-04 ***
## former smoker - current smoker == 0  0.4463     0.2663   1.676    0.214
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

The results indicates that there is an association between the smoking status and the risk of MI. The p-value for this association is defined as the minimun of the p-values of all pairwise comparisons, hence here p-value <0.0001. We can further conclude that there are two significant differences; between the risk of MI of:

- women who are former smokers and those who have never smoked (p-value <0.0001)
- women who are current smokers and those who have never smoked (p-value <0.0001)

The results do not show a significant difference between the risk of MI of women who are former smokers and those who are current smokers (p-value=0.214).

To obtain the estimated odds ratios with adjusted 95% confidence intervals, we use the following code. The idea is that by default the results are presented for the parameters, which are the logarithm of odds ratios. Hence, to get the results for the odds ratios we take the exponential.

```
exp(confint(Res1)$confint)
```

```
##                                Estimate       lwr       upr
## current smoker - never smoked  4.258824 2.3445512  7.736055
## former smoker - never smoked   6.654412 3.4909999 12.684388
## former smoker - current smoker 1.562500 0.8373958  2.915475
## attr(,"conf.level")
## [1] 0.95
## attr(,"calpha")
## [1] 2.342394
```

Hence we can update our conclusions as follows. The results indicate that there is a significance difference in the risk of MI between:

- women who are former smokers and those who have never smoked (OR=4.26, 95% CI=[2.34;7.74], p-value <0.0001)
- women who are current smokers and those who have never smoked (OR=6.65, 95% CI=[3.49;12.7],p-value <0.0001)

Question 3

We now fit a "simple" (univariate) logistic model to investigate the association between myocardial infarction (**mi**) and use of oral contraceptives (**oc**).

```
fit2 <- glm(mi~oc, data=MI, family=binomial)
summary(fit2)
```

```
## 
## Call:
## glm(formula = mi ~ oc, family = binomial, data = MI)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.2814  -0.5672  -0.5672   1.0769   1.9527  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -1.7457     0.1782  -9.798   <2e-16 ***
## oc            1.9868     0.2281   8.710   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 570.66  on 448  degrees of freedom
## Residual deviance: 483.66  on 447  degrees of freedom
## AIC: 487.66
## 
## Number of Fisher Scoring iterations: 4
```

```
publish(fit2)
```

```
##  Variable Units OddsRatio        CI.95 p-value
##        oc            7.29 [4.66;11.40] < 1e-04
```

The results indicate that the risk of MI is significantly higher for women who use oral contraceptives than for those who do not (OR=7.29, 95% CI=[4.66;11.40], p-value<0.0001).

Question 4

We now compute a frequency (3 by 2) table to compare the proportions of women who are current smokers, former smokers or who have never smoked, among those who use oral contraceptives and those who do not use them.

First we look at the counts.

```
Tab1 <- table(MI$oc,MI$Smoke)
Tab1
```

```
## 
##     never smoked current smoker former smoker
##   0          140             72            37
##   1           75             63            62
```

We now look at the proportions. We use the **margin=1** options to indicate that we want the proportions by use of oral contaceptives, i.e. by line (we would use **margin=2** to obtained the proportions by column).

```
prop.table(Tab1,margin=1)
```

```
## 
##     never smoked current smoker former smoker
##   0    0.5622490      0.2891566     0.1485944
##   1    0.3750000      0.3150000     0.3100000
```

We can see that among users of oral contraceptives there are less women who have never smoked (37.5%) than among those who do not use them (56.2%). This is interesting because it means that when we compared women who use contraceptives to those wo do not at the previous question we were **implicitely** comparing:

- women who use oral contraceptives **and** have "often" a current or past history of smoking (62%)
- to women who do not use oral contraceptives **and** have "less often" a current or past history of smoking (44%).

As the two groups are different with respect to both use of oral contraceptives and smoking status, we conclude that from the previous results it seems difficult to know what is the "reason" for the difference in risks of MI that we observe between the two groups. It could be due to either use of oral contraceptives or smoking status/history or both (or in fact another difference between the two groups). As it is known that smoking influences the risk of MI on its own, this makes the results to the previous question not very informative to answer our research question.

We now estimate a (multiple) logistic model to model the risk of myocardial infarction (**mi**) using the two variables corresponding to smoking status (**Smoke**) and use oral contraceptives (**oc**). Here we are asked to not model an interaction.

```
fit3 <- glm(mi ~ oc + Smoke, data=MI, family=binomial)
publish(fit3)
```

```
##  Variable          Units OddsRatio        CI.95 p-value
##        oc                     6.71 [4.18;10.77]  <1e-04
##     Smoke   never smoked       Ref                     
##           current smoker      4.34  [2.51;7.51]  <1e-04
##            former smoker      5.32  [2.96;9.56]  <1e-04
```

We can have the following interpretations. From this model, we estimate that, when comparing two women:

- one uses oral contraceptives, the other does not, both have the same smoking status/history (whatever it is), the risk of MI of the user of oral contraceptives is significantly higher (OR=6.71, 95% CI=[4.18;11.77], p-value<0.0001).
- one is a current smoker, the other has never smoked, both have the same use of oral contaceptives (whatever it is, either they both use them or both do not use them), the risk of MI of the current smoker is significantly higher (OR=4.34, 95% CI=[2.51;7.51], p-value<0.0001).
- one is a former smoker, the other has never smoked, both have the same use of oral contaceptives (whatever it is, either they both use them or both do not use them), the risk of MI of the former smoker is significantly higher (OR=5.32, 95% CI=[2.96;9.56], p-value<0.0001).

According to the first item above, the odds ratio which compares the risk of users of oral contraceptives to that of non users is **the same whatever the smoking status/history**. This seems to be inconsistent with what previous studies suggested, i.e. "*oral contraceptives could increase the risk of MI differently for smokers, non-smokers and former smokers*". What is important to notice is that this inconsistency between previous results and ours is not driven by the data but, instead, by the modeling assumption of no interaction. In other words, our modeling assumption does not allow us to estimate a different association between MI and use of oral contraceptive for women who smoke, do not smoke or have never smoked, although we should, according to previous studies.

Question 5

We now estimate a (multiple) logistic model to model the risk of myocardial infarction using the two variables corresponding to smoking status (**Smoke**) and oral contraceptives (**oc**) and we model an **interaction** between the two variables.

```r
fit4 <- glm(mi ~ oc * Smoke, data=MI, family=binomial)
summary(fit4)
```

```
##
## Call:
## glm(formula = mi ~ oc * Smoke, family = binomial, data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8597  -0.7828  -0.4590   0.6250   2.1460
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)            -2.19722    0.28169  -7.800 6.19e-15 ***
## oc                      1.18562    0.38410   3.087 0.002024 **
## Smokecurrent smoker     1.17137    0.38840   3.016 0.002562 **
## Smokeformer smoker      0.08701    0.59970   0.145 0.884638
## oc:Smokecurrent smoker  0.46276    0.53747   0.861 0.389244
## oc:Smokeformer smoker   2.45852    0.73371   3.351 0.000806 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 570.66  on 448  degrees of freedom
## Residual deviance: 425.94  on 443  degrees of freedom
## AIC: 437.94
##
## Number of Fisher Scoring iterations: 4
```

We further ask for the "formatted" results using the **publish()** function. This facilitates the interpretation of the results, especially with such a model with an interaction.

```r
publish(fit4)
```

```
##                   Variable Units OddsRatio          CI.95   p-value
##    oc: Smoke(never smoked)            3.27    [1.54;6.95]  0.002024
##  oc: Smoke(current smoker)            5.20   [2.49;10.86]   < 1e-04
##   oc: Smoke(former smoker)           38.25 [11.23;130.24]   < 1e-04
```

We can have the following interpretations. From this model, we estimate that when comparing the risk of MI of a user of oral contraceptives to that of a woman who does not use oral contraceptives:

- when both have **never smoked**, the risk of MI of the user of oral contraceptives is significantly higher (OR=3.27, 95% CI=[1.54;6.95], p-value=0.002).

- when both are **current smokers**, the risk of MI of the user of oral contraceptives is significantly higher (OR=5.20, 95% CI=[2.49;10.86], p-value<0.0001).
- when both are **former smokers**, the risk of MI of the user of oral contraceptives is significantly higher (OR=38.25, 95% CI=[11.23;130.24], p-value<0.0001).

Note that these three odds ratio can be computed "by hand" from the estimated parameters shown by the **summary** function as follows:

```r
exp(1.18562)
```

```
## [1] 3.272715
```

```r
exp(1.18562 + 0.46276)
```

```
## [1] 5.198551
```

```r
exp(1.18562 + 2.45852)
```

```
## [1] 38.24986
```

Hence, the exponential of the parameter at the lines **oc** in the output of the summary function (i.e. 1.18562) corresponds to the log of the odds ratio to for use of oral contraceptive for the **reference** smoking group, i.e. those who have never smoked. Hence the confidence interval for the OR could also be obtained as follows:

```r
exp(confint.default(fit4)["oc",])
```

```
##    2.5 %   97.5 %
## 1.541567 6.947960
```

The other confidence intervals could be obtained similarly after changing the reference level for the variable "Smoke". For instance:

```r
MI$Smokeb <- relevel(MI$Smoke,ref="current smoker")
fit4b <- glm(mi ~ oc * Smokeb, data=MI, family=binomial)
exp(coef(fit4b)["oc"])             # odds ratio
```

```
##       oc
## 5.198565
```

```r
exp(confint.default(fit4b)["oc",]) # confidence interval for the odds ratio
```

```
##     2.5 %    97.5 %
##  2.488099 10.861735
```

We can see that the results suggest that use of oral contraceptives is differently associated with MI depending on the smoking status. Indeed, the confidence intervals are quite different for the women who are current smokers, former smokers and those who have never smoked. Furthermore, in the output of the summary function we can read:

- `oc:Smokeformer smoker 2.45852 0.73371 3.351 0.000806 ***`

This suggests that the association between MI and oral contraceptives is significantly different among women who have never smoked and among woman who are former smokers (p-value=0.000806). The parameter estimate 2.45852 represent the log of the ratio of the two odds ratios, i.e.:

```
log(38.25/3.27)
```

```
## [1] 2.459354
```

Note: the above results does not match "exactly" the parameter estimate 2.45852 here only because we have used the estimates of the two odds ratios 38.25 and 3.27 rounded at the second digits (not their exact estimated value).

Question 6

We now make some boxplots to compare the distribution of ages in the six groups of women defined by all possible combinations of smoking status and use of oral contraceptives.

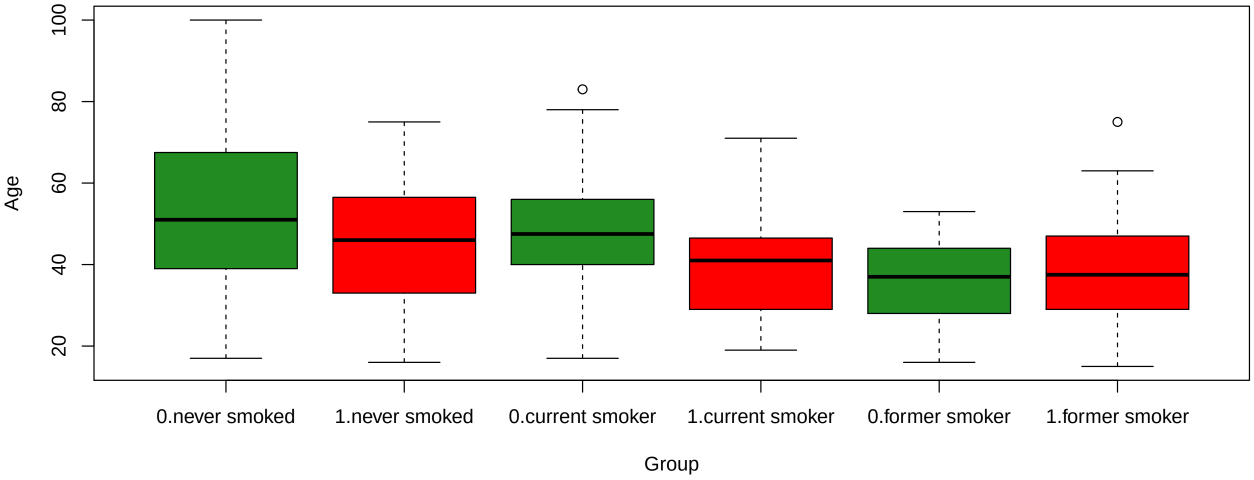
We first make a new variables to indicate in which of the six groups each woman belongs (and look at how many women there are in each group).

```
MI$group <- interaction(MI$oc,MI$Smoke)
table(MI$group)
```

```
##
##   0.never smoked   1.never smoked 0.current smoker 1.current smoker
##              140               75               72               63
##  0.former smoker  1.former smoker
##               37               62
```

We can now easily make the boxplots. We use the red color for users of oral contraceptives and green for the others.

```
boxplot(MI$age~interaction(MI$oc,MI$Smoke),
        xlab="Group",
        ylab="Age",
        col=rep(c("forestgreen","red"),3))
```



From the boxplot we can see that when we compare women who use oral contraceptives to those who do not, both having the same smoking status, the two groups are not necessarily similar with respect to age. For instance, current smokers who use oral contraceptives seem to be often younger than current smokers who do not use oral contraceptives. Hence, although we found at the previous question that:

- the risk of MI of the user of oral contraceptives is significantly higher to that of women who do not use contraceptives, among current smokers (OR=5.20, 95% CI=[2.49;10.86], p-value<0.0001);

from a purely statistical point of view, it is difficult to say whether the difference in risk of MI that we see between these two groups comes from the use of oral contraceptives or from the difference in age (or maybe another difference between the two groups). However, we can reasonably think that if the two compared groups had been similar with respect to age, we would have seen an even larger difference in the risk of MI. Indeed, those who use contraceptives are younger and it is known that aging (in its own) increases the risk of MI.

Part II

We now proceed to the main analysis of the data, which aims to shed light on the research question and lead to the main conclusions. We will estimate a model similar to that of question 5 but we will additionally adjust on age. Hence, we will be able to compare women who use contraceptives to those who do not, among women who are similar with respect to smoking and age.

Question 7

We first create add and to the data a categorical variable **AgeGroup**, which is a categorical variable for age, with three groups 15-39, 40-55 and 55 or above. We further use the **table()** function to read the number of observations in each group.

```
MI$AgeGroup <- cut(MI$age,breaks=c(15,40,55,100),include.lowest=TRUE)
table(MI$AgeGroup)
```

```
##
##  [15,40]  (40,55] (55,100]
##      175      160      114
```

Question 8

We now estimate a (multiple) logistic model to model the risk of myocardial infarction (**mi**) using the three variables corresponding to smoking status (**Smoke**), use of oral contraceptives (**oc**) and age group (**AgeGroup**). We model an interaction between smoking status (**Smoke**) and use of oral contraceptives (**oc**)

```
fit5 <- glm(mi ~ oc * Smoke + AgeGroup, data=MI, family=binomial)
summary(fit5)
```

```
##
## Call:
## glm(formula = mi ~ oc * Smoke + AgeGroup, family = binomial,
##     data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1086  -0.5880  -0.4108   0.7315   2.3092
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -3.3564     0.4090  -8.207 2.27e-16 ***
## oc                        1.4792     0.4041   3.660 0.000252 ***
## Smokecurrent smoker       1.3792     0.4064   3.393 0.000690 ***
## Smokeformer smoker        0.7623     0.6388   1.193 0.232722
## AgeGroup(40,55]           0.9267     0.3178   2.916 0.003544 **
## AgeGroup(55,100]          1.6572     0.3727   4.446 8.75e-06 ***
## oc:Smokecurrent smoker    0.6068     0.5604   1.083 0.278952
## oc:Smokeformer smoker     2.2968     0.7581   3.030 0.002449 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 570.66  on 448  degrees of freedom
## Residual deviance: 403.14  on 441  degrees of freedom
## AIC: 419.14
```

```
##
## Number of Fisher Scoring iterations: 5
```

```r
publish(fit5)
```

```
##                    Variable    Units OddsRatio           CI.95    p-value
##                    AgeGroup  [15,40]       Ref
##                              (40,55]      2.53     [1.36;4.71]   0.003544
##                             (55,100]      5.24    [2.53;10.89]    < 1e-04
##    oc: Smoke(never smoked)                4.39     [1.99;9.69]   0.000252
##  oc: Smoke(current smoker)                8.05    [3.59;18.04]    < 1e-04
##   oc: Smoke(former smoker)               43.64 [12.37;153.91]    < 1e-04
```

This model seems reasonable for the main analysis for the following reasons:

- it enables us to study the association between use of oral contraceptives and MI.
- we adjust on age group and smoking status to compare the risk of MI between women who use and those who do not use contraceptives, who are otherwise similar with respect to smoking status and age (group). This seems sensible because 1) previous studies suggest that both smoking status and age influence the risk of MI and 2) the study design (observational study) does not ensure that the group of women who use contraceptive is "similar" to that of those who do not, in terms of age and smoking status.
- as we use an interaction between smoking and use of contraceptives, we do not assume that the association between use of oral contraceptives and MI is the same in all smoking groups (i.e. we do not force the model to estimate the same association between contraceptives and MI in all smoking groups).

From this model, we estimate that when comparing the risk of MI of a user of oral contraceptives to that of a woman who does not use oral contraceptives:

- when both have **never smoked** and belong to the **same age group** (whatever it is), the risk of MI of the user of oral contraceptives is significantly higher (OR=4.39, 95% CI=[1.99;9.69], p-value=0.002).
- when both are **current smokers** and belong to the **same age group** (whatever it is), the risk of MI of the user of oral contraceptives is significantly higher (OR=8.05, 95% CI=[3.59;18.04], p-value<0.0001).
- when both are **former smokers** and belong to the **same age group** (whatever it is), the risk of MI of the user of oral contraceptives is significantly higher (OR=43.64, 95% CI=[12.37;153.91], p-value<0.0001).

There is a significant association in each smoking group, but it seems that its strength differs from one smoking group to another. For instance, the output of the summary function shows that the odds ratios modeling the association between MI and use of contraceptives seem different for former smokers and those who have never smoked (p-value=0.002449). The difference in odds ratio is not significant between current smokers and those who have never smoked (p-value=0.278952). Note that the third comparison of odds ratios, between current and former smokers is not shown in the output. To get it, you can simply refit the model after changing the reference level as shown below and see that the difference is again significant (p-value=0.02419).

```
MI$Smokeb <- relevel(MI$Smoke,ref="current smoker")
fit5b <- glm(mi ~ oc * Smokeb + AgeGroup, data=MI, family=binomial)
summary(fit5b)
```

```
##
## Call:
## glm(formula = mi ~ oc * Smokeb + AgeGroup, family = binomial,
##     data = MI)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1086  -0.5880  -0.4108   0.7315   2.3092
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.9772     0.3726  -5.306 1.12e-07 ***
## oc                        2.0859     0.4115   5.070 3.99e-07 ***
## Smokebnever smoked       -1.3792     0.4064  -3.393  0.00069 ***
## Smokebformer smoker      -0.6169     0.6139  -1.005  0.31491
## AgeGroup(40,55]           0.9267     0.3178   2.916  0.00354 **
## AgeGroup(55,100]          1.6572     0.3727   4.446 8.75e-06 ***
## oc:Smokebnever smoked    -0.6068     0.5604  -1.083  0.27895
## oc:Smokebformer smoker    1.6901     0.7498   2.254  0.02419 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 570.66  on 448  degrees of freedom
## Residual deviance: 403.14  on 441  degrees of freedom
## AIC: 419.14
##
## Number of Fisher Scoring iterations: 5
```

Part III

Question 9

We now perform a "sensitivity analysis" by changing the way the variable age enters the multiple logistic regression model. We now use the variable age as a continuous variable (assuming a linear effect of age). To make the interpretation somehow "easier", we create the variable **age20**, which is the variable **age** divided by 20.

```
MI$age20 <- MI$age/20
fit6 <- glm(mi ~ oc * Smoke + age20, data=MI, family=binomial)
publish(fit6)
```

```
##                    Variable Units OddsRatio          CI.95 p-value
##                       age20            2.98    [2.04;4.35] < 1e-04
##    oc: Smoke(never smoked)             6.50   [2.77;15.24] < 1e-04
##  oc: Smoke(current smoker)             9.69   [4.27;22.00] < 1e-04
##   oc: Smoke(former smoker)            46.02 [13.10;161.64] < 1e-04
```

We can see that the main results, that is, the results for the association between use of contraceptives and risk of MI, are relatively unchanged. Of course the estimated values are a bit different, but the confidence intervals that we obtain are very similar (in the sense that they overlap a lot) with this model and the main model. Hence the conclusions that we draw from this model or the main model are similar. The small differences between the results of the two models are irrelevant for the clinical interpretation of the results. This is somehow reassuring: the results do not heavily depend on the (somehow) arbitrary way we have modelled the effect of age.