Lab 3 - Logistic Regression and Healthcare Data

Jerry Zikun Chen

2019-10-29

Part 1 - Kidney Stone Treatement Data

a)

First, we recreate the dataset from the paper. Among the patients that have kidney stones with mean diameter less than 2 centimeters, there are in total 87 patients received open surgery (open_surg), and 270 patients received Percutaneous nephrolithotomy ESWL (percut) treatment. There are 263 open surgeries and 80 percutaneous treatments performed on patients that have kidney stones with mean diameter larger than or equal to 2 centimeters. Successes are encoded as 1’s and failures are encoded as 0’s.

```r
library(tidyverse)
library(broom)
library(plyr)
library(dplyr)
library(plotROC)
options(digits=5)
```

```r
# Create dataset from the paper
group <- c(rep('<2', 87), rep('<2', 270), rep('>=2', 263), rep('>=2', 80))
proc <- c(rep('open_surg', 87),  rep('percut', 270), rep('open_surg', 263), rep('percut'
, 80))
success <- as.integer(c(rep(1, 81), rep(0, 6),
                        rep(1, 234), rep(0, 36),
                        rep(1, 192), rep(0, 71),
                        rep(1, 55), rep(0, 25)))

kidney_df <- as_tibble(data.frame(group, proc, success))
head(kidney_df)
```

```
## # A tibble: 6 x 3
##   group proc      success
##   <fct> <fct>       <int>
## 1 <2    open_surg       1
## 2 <2    open_surg       1
## 3 <2    open_surg       1
## 4 <2    open_surg       1
## 5 <2    open_surg       1
## 6 <2    open_surg       1
```

Logistic Regression without Seperation of Kidney Stone Size

```
table(kidney_df$proc, kidney_df$success)
```

```
## 
##                0   1
##   open_surg  77 273
##   percut     61 289
```

By the contingency table above, we can see that open sugery group have a success rate of \(273/(273+77) = 0.78\), lower than the percut group which has a success rate of \(289/(289+61) = 0.826\). The odds ratio of sucess in percut vs. open surgery group is \((289/61)/(273/77) = 1.337\), which means percut operations are \(33.7\%\) more likely to be successful on kidney stone patients than open sugeries based on the dataset.

```
lr_full <- glm(success ~ proc, data = kidney_df, family = binomial)
tidy(lr_full)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.27      0.129      9.81 1.03e-22
## 2 procpercut     0.290     0.191      1.52 1.29e- 1
```

```
exp(cbind("odds ratio" = coef(lr_full), confint(lr_full)))
```

```
## Waiting for profiling to be done...
```

```
##             odds ratio  2.5 % 97.5 %
## (Intercept)     3.5455 2.7687 4.5948
## procpercut      1.3363 0.9200 1.9479
```

We fit the logistic regression model (\(\pi\) is probability of success and categorical variable \(x\) is the procedure type): \[log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1x\] As wee can see from the above summary, the p-value for \(\beta_1\) is greater than 0.05. We are not statistically confident to conclude that \(\beta_1\) has the value of 0.29, meaning that we cannot say that if we change the procedure from an open surgery to a percut, then we will have in \(exp(0.29) = 1.337\) change in odds of success. Note that \(exp(0.29)\) matches the odds ratio from the contingency table. The confidence interval of log odds includes the estimate \(1\) for \(\beta_1\), meaning that there is no difference between the odds of success of the two operation.

b)

Next, we compare results in two seperate patient groups with kidney stones with diameter greater than or equal to 2 centimeters and smaller than 2 centimeters respectively. We produce contigency tables and fit logistic regression models to these two groups.

Logistic Regression for >=2 Group

```
kidney_large <- filter(kidney_df, group == ">=2")
table(kidney_large$proc, kidney_large$success)
```

```
##
##                0   1
##   open_surg   71 192
##   percut      25  55
```

By the contingency table of \(>=2\) group, we can see that open sugery group have a success rate of \(192/(192+71) = 0.73\), higher than the percut group, which has a success rate of \(55/(55+25) = 0.6875\). The odds ratio of success in percut vs. open surgery group is \((55/25)/(192/71) = 0.8135\). Therefore, according to the table alone, percut operations are \(19\%\) less likely to be successful on patients with larger kidney stones.

```
lr_large <- glm(success ~ proc, data = kidney_large, family = binomial)
tidy(lr_large)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.995     0.139     7.16  7.94e-13
## 2 procpercut    -0.206     0.278    -0.741 4.58e- 1
```

```
exp(cbind("odds ratio" = coef(lr_large), confint(lr_large)))
```

```
## Waiting for profiling to be done...
```

```
##             odds ratio   2.5 % 97.5 %
## (Intercept)    2.70423 2.07058 3.5723
## procpercut     0.81354 0.47495 1.4194
```

The p-values for \(\beta_1\) is not statistically significant so we cannot say that for patients with larger kidney stones, if we change the procedure from an open surgery to a percut, then we will have significant changes in odds of success. This inconclusive result is also confirmed by the confidence interval where \(1\) is included in the estimate of the slope.

Logistic Regression for <2 Group

```
kidney_small <- filter(kidney_df, group == "<2")
table(kidney_small$proc, kidney_small$success)
```

```
##
##                0   1
##   open_surg    6  81
##   percut      36 234
```

By the contingency table of \(>=2\) group, we can see that open sugeries have a success rate of \(81/(81+6) = 0.93\), higher than the percut group which has a success rate of \(243/(243+36) = 0.87\). The odds ratio of success in percut vs. open surgery group is \((234/36)/(81/6) = 0.48\). Therefore, it suggests that compared to open sugeries, percut operations are \(48\%\) less likely to be successful on patients with smaller kidney stones.

```
lr_small <- glm(success ~ proc, data = kidney_small, family = binomial)
tidy(lr_small)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    2.60      0.423      6.15 7.67e-10
## 2 procpercut    -0.731     0.459     -1.59 1.12e- 1
```

```
exp(cbind("odds ratio" = coef(lr_small), confint(lr_small)))
```

```
## Waiting for profiling to be done...
```

```
##             odds ratio   2.5 % 97.5 %
## (Intercept)   13.50000 6.41736 34.718
## procpercut     0.48148 0.17726  1.107
```

Similar to the other group, the p-value for the slope is again statistically insignificant and we cannot be confident about these estimates. Confidence interval for the slope includes an estimate of \(1\).

Confusion Matrix and ROC curve

Next, we analyze confusion matrices and ROC curves for the two groups.

```
pred_probs <- lr_large %>% predict(type = "response")
pred_class  <- ifelse(pred_probs >= 0.70, "positive", "failure")
confusion <- table(kidney_large$success, pred_class)

tp <- confusion[2,2]
fp <- confusion[1,2]
fn <- confusion[2,1]
tn <- confusion[1,1]
recall <- tp / (fn + tp)
precision <- tp /(tp + fp)
accuracy <- (tp + tn) / (tp + tn + fp + fn)
f_score <- (2*precision*recall)/(recall+precision)
confusion
```

```
##    pred_class
##     failure positive
##   0      25       71
##   1      55      192
```

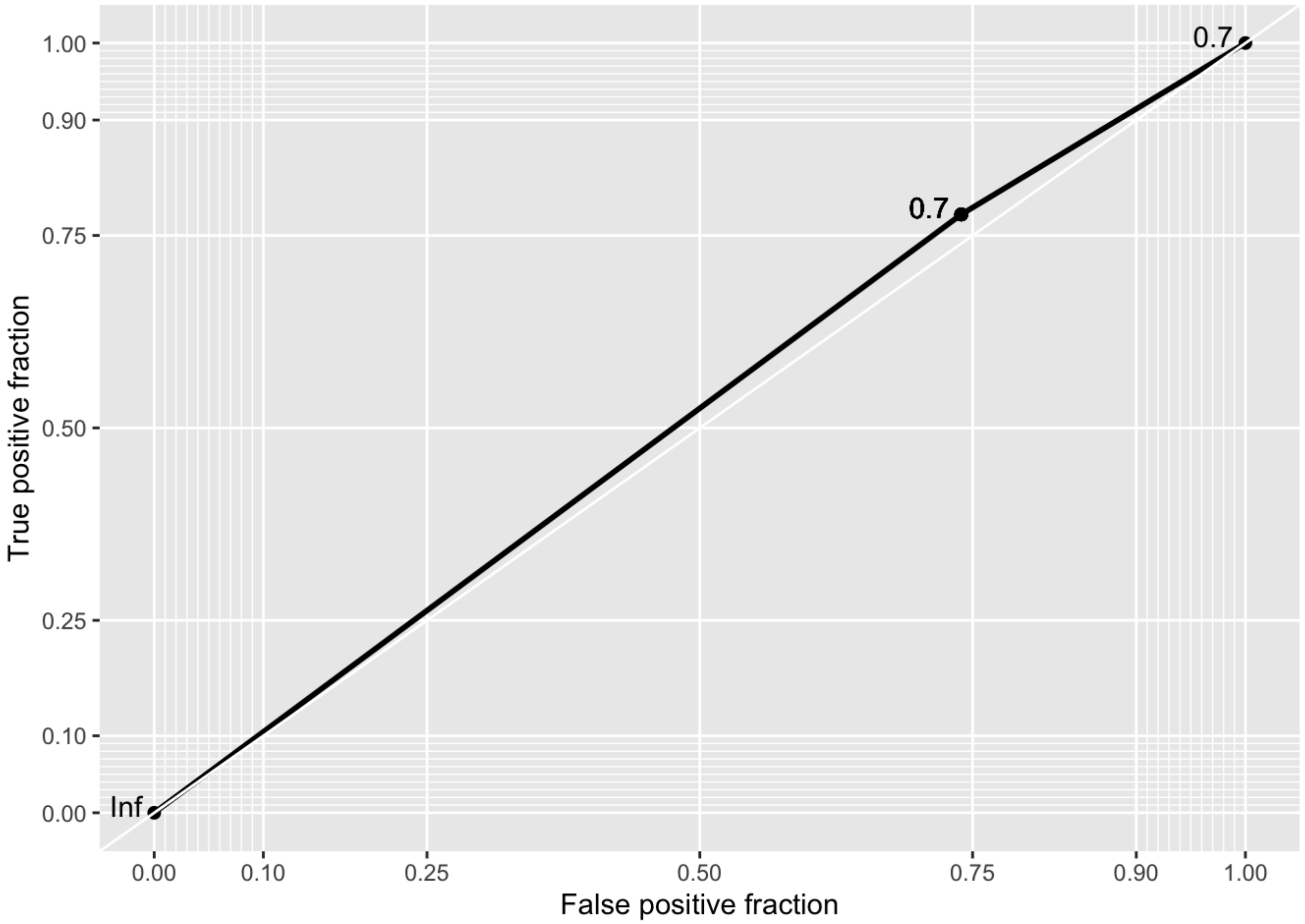
```
sprintf("accuracy: %.4f", accuracy)
```

```
## [1] "accuracy: 0.6327"
```

```
sprintf("F-score: %.4f", f_score)
```

```
## [1] "F-score: 0.7529"
```

```
tibble(pred = pred_probs, obs = kidney_large$success) %>%
  ggplot(aes(d = obs, m = pred)) + geom_roc() + style_roc(theme = theme_gray)
```



```
pred_probs <- lr_small %>% predict(type = "response")
pred_class  <- ifelse(pred_probs >= 0.9, "positive", "negative")
confusion <- table(kidney_small$success, pred_class)

tp <- confusion[2,2]
fp <- confusion[1,2]
fn <- confusion[2,1]
tn <- confusion[1,1]
recall <- tp / (fn + tp)
precision <- tp /(tp + fp)
accuracy <- (tp + tn) / (tp + tn + fp + fn)
f_score <- (2*precision*recall)/(recall+precision)
confusion
```

```
##    pred_class
##     negative positive
##   0       36        6
##   1      234       81
```

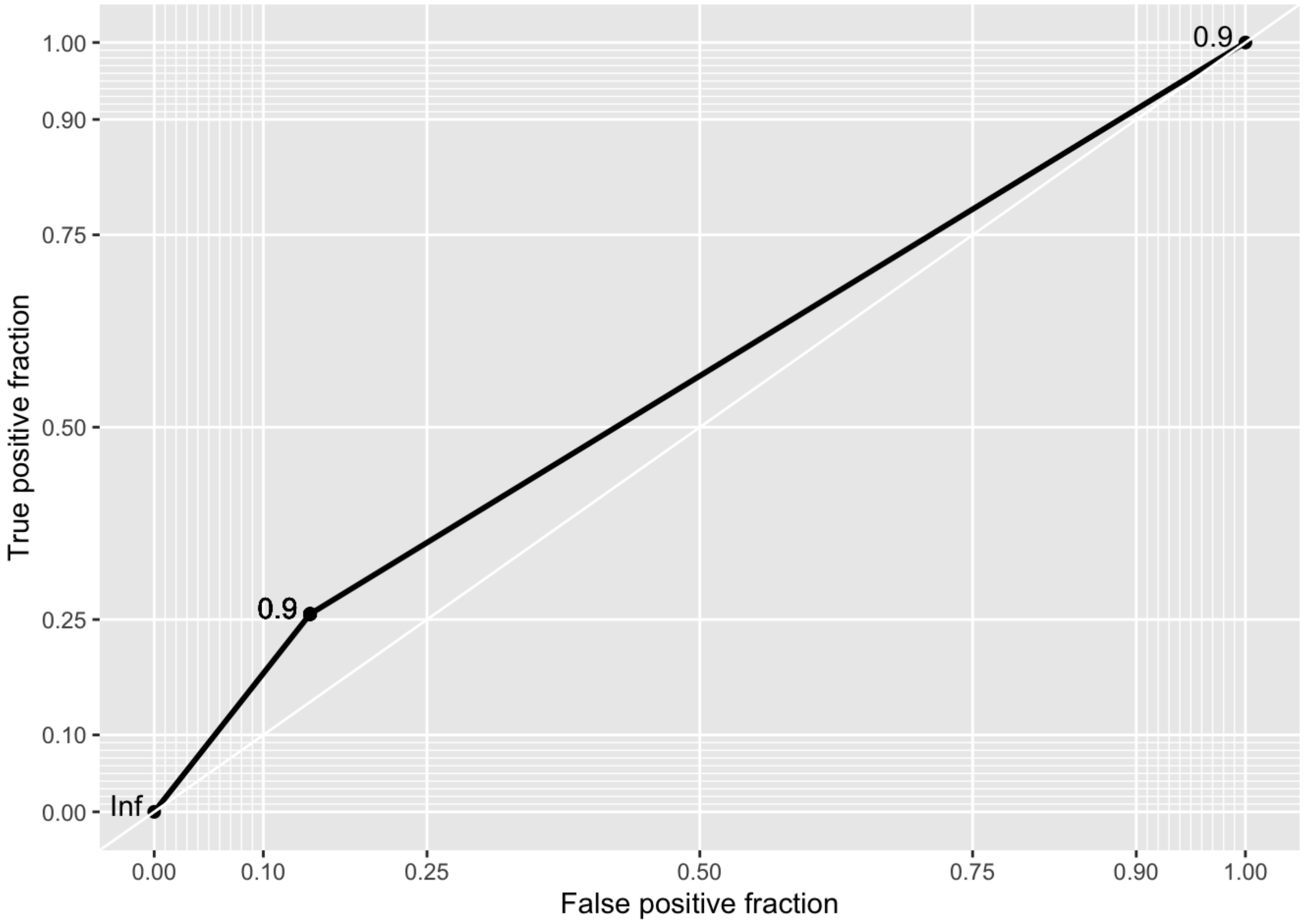
```
sprintf("accuracy: %.4f", accuracy)
```

```
## [1] "accuracy: 0.3277"
```

```
sprintf("F-score: %.4f", f_score)
```

```
## [1] "F-score: 0.4030"
```

```
tibble(pred = pred_probs, obs = kidney_small$success) %>%
  ggplot(aes(d = obs, m = pred)) + geom_roc() + style_roc(theme = theme_gray)
```



We can see that neither the model for the smaller nor the larger group achieved good accuracy when the best thresholds were selected. The ROC plots are only slight better then random guesses.

Probit Regression

We further investigate the dataset with probit link functions below:

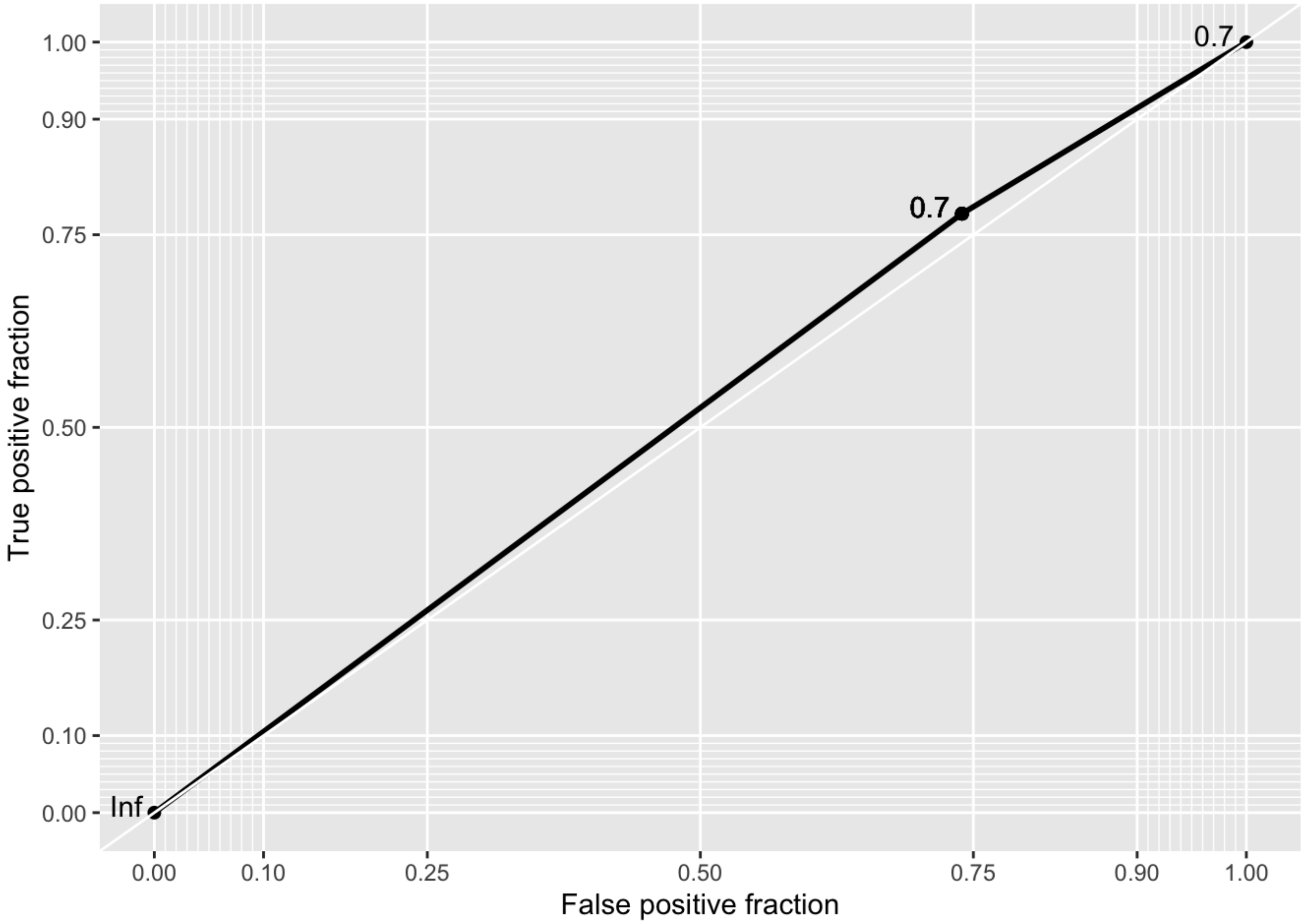
```
# Probit Model for >=2 group
probit_large <- glm(success ~ proc, family = binomial(link = "probit"), data = kidney_la
rge)
tidy(probit_large)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    0.613    0.0828     7.40  1.33e-13
## 2 procpercut    -0.124    0.168     -0.738 4.60e- 1
```

```
pred_probs <- probit_large %>% predict(type = "response")
pred_class  <- ifelse(pred_probs >= 0.7, "succ", "fail")
table(pred_class, kidney_large$success)
```

```
##
## pred_class   0   1
##       fail  25  55
##       succ  71 192
```

```
tibble(pred = pred_probs, obs = kidney_large$success) %>%
  ggplot(aes(d = obs, m = pred)) + geom_roc() + style_roc(theme = theme_gray)
```



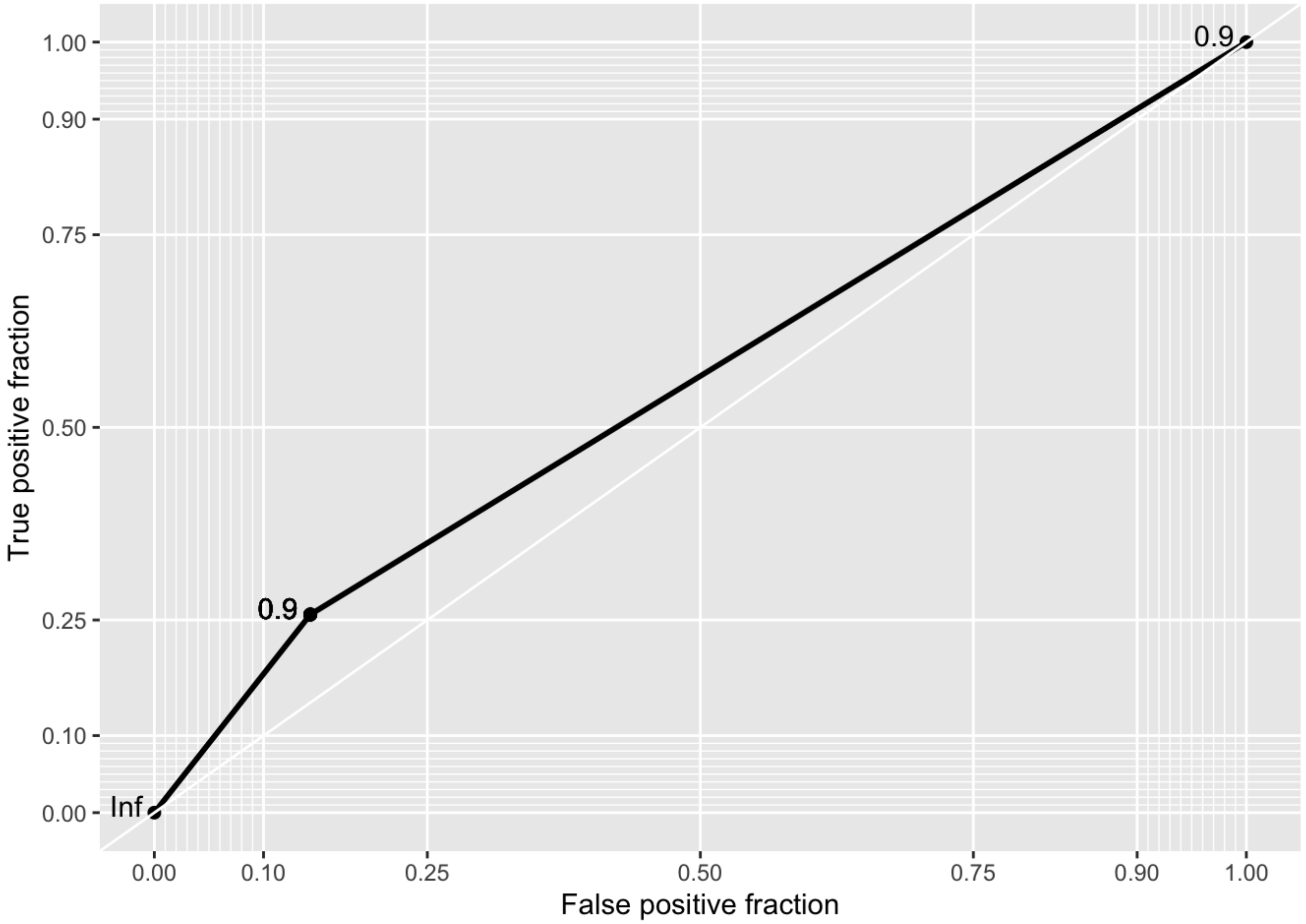
```
# Probit Model for <2 group
probit_small <- glm(success ~ proc, family = binomial(link = "probit"), data = kidney_sm
all)
tidy(probit_small)
```

```
## # A tibble: 2 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.48      0.205      7.25 4.21e-13
## 2 procpercut    -0.373     0.226     -1.65 9.92e- 2
```

```
pred_probs <- probit_small %>% predict(type = "response")
pred_class  <- ifelse(pred_probs >= 0.9, "succ", "fail")
table(pred_class, kidney_small$success)
```

```
##
## pred_class   0   1
##       fail  36 234
##       succ   6  81
```

```
tibble(pred = pred_probs, obs = kidney_small$success) %>%
  ggplot(aes(d = obs, m = pred)) + geom_roc() + style_roc(theme = theme_gray)
```



```
# Full Probit Model for all patients
probit_full <- glm(success ~ proc + group, family = binomial(link = "probit"), data = ki
dney_df)
tidy(probit_full)
```

```
## # A tibble: 3 x 5
##   term        estimate std.error statistic  p.value
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)    1.36      0.136      9.99 1.60e-23
## 2 procpercut    -0.215     0.132     -1.63 1.02e- 1
## 3 group>=2      -0.721     0.134     -5.39 6.89e- 8
```

```
pred_probs <- probit_full %>% predict(type = "response")
pred_class  <- ifelse(pred_probs >= 0.7, "succ", "fail")
confusion <- table(pred_class, kidney_df$success)

tp <- confusion[2,2]
fp <- confusion[1,2]
fn <- confusion[2,1]
tn <- confusion[1,1]
recall <- tp / (fn + tp)
precision <- tp /(tp + fp)
accuracy <- (tp + tn) / (tp + tn + fp + fn)
f_score <- (2*precision*recall)/(recall+precision)
confusion
```

```
## 
## pred_class   0   1
##       fail  25  55
##       succ 113 507
```

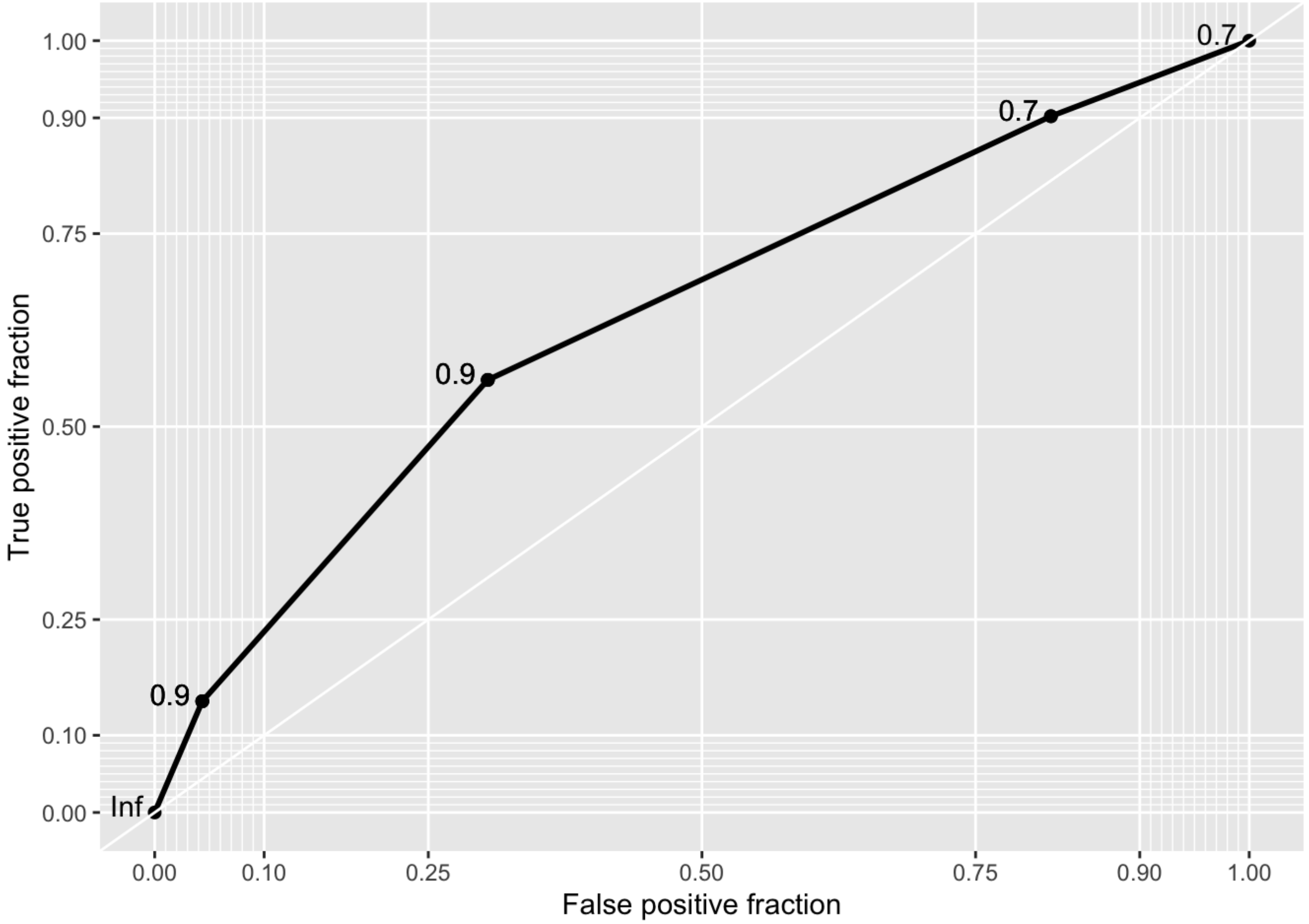
```
sprintf("accuracy: %.4f", accuracy)
```

```
## [1] "accuracy: 0.7600"
```

```
sprintf("F-score: %.4f", f_score)
```

```
## [1] "F-score: 0.8579"
```

```
tibble(pred = pred_probs, obs = kidney_df$success) %>%
  ggplot(aes(d = obs, m = pred)) + geom_roc() + style_roc(theme = theme_gray)
```



From the probit regression models, we can see that the slopes became statistically significant. Amount the smaller group, the probit model shows that percut operations can raise the probability of success by \(\Phi(-0.124) = 0.45\). In the second model for \(<2\) group, there is a \(\Phi(-0.373) = 0.35\) increase in the probability of success for the larger group. Despite the lack of statistical significance in the coefficient for the operation type, the full probit model achieves good prediction results with accuracy of \(0.76\) and a F-score of \(0.86\).

c) Conclusion

From the logistic regression analysis, we can conclude that it is neccesary to treat patients with kidney stones of larger and smaller size differently based on the disproportionate operation assignment. They were not randomly assigned to patients. Most of patients with larger kidney stones were treated with open surgeries and vice versa. For the \(<2\) group, the result is not clear since we did not achieve a statistically significant estimate of the log odds ratio. Even though the open surgeries has a higher rate of success empirically, most patients in this group are treated with percut so there can be more variabilities in the percut results. This assignment bias is the opposite for the \(>=2\) group. So there is no meaningful conclusions to be drawn about which operation is better for neither groups. Prediction results for both groups are poor according to confusion matrices and ROC curves. We do not have a clear-cut conclusion for neither group based on logistic regression models.

On the other hand, the probit regression models provides a better story. We estimate that percut operation is better than open surgeries. This is based on the statistically significant estimates that we have achieved in both groups. Specifically, we expect a \(0.45\) and \(0.35\) increases in probabilities of operation success for the larger and smaller groups respectively. Furthermore, the full probit model consist of both the procedure type and kidney stone size achieve as input achieves better prediction results based on the confusion matrices and its ROC curve. However, it is worth noting that when patients are seperated into two groups, the logistic regression predictions are not as good. When looked together, the coefficient of operation type becomes less significant.

Part 2 - Risk of Cardiovascular Disease among Osteoarthritis Patients (Statistical Society of Canada)

a)

To answer questions 1,2, and 4 from the website: https://ssc.ca/en/case-study/case-study-2-risk-cardiovascular-disease-among-osteoarthritis-patients (https://ssc.ca/en/case-study/case-study-2-risk-cardiovascular-disease-among-osteoarthritis-patients) We first load and clean the datasets. Datasets are loaded from Rdata files in the data folder.

```
load("data/cchs11.Rdata")
load("data/cchs21.Rdata")
load("data/cchs31.Rdata")
```

```
# select 12 out of 23 columns
col_names_11 <- c("CCCA_121", "CCCA_05A", "DHHAGAGE", "DHHA_SEX", "SDCAGRAC", "EDUADR04"
, "INCAGHH", "HWTAGBMI",
                  "TWDA_5", "SMKADSTY", "ALCADTYP", "CCCA_071", "CCCA_101", "PACADPAI",
"GEOAGPRV", "SDCAGRES", "DHHAGMS")
col_names_21 <- c("CCCC_121", "CCCC_05A", "DHHCGAGE", "DHHC_SEX", "SDCCGRAC", "EDUCDR04"
, "INCCGHH", "HWTCGBMI",
                  "HCUC_1AA", "SMKCDSTY", "ALCCDTYP", "CCCC_071", "CCCC_101", "PACCDPAI"
, "GEOCGPRV", "SDCCGRES", "DHHCGMS")
col_names_31 <- c("CCCE_121", "CCCE_05A", "DHHEGAGE", "DHHE_SEX", "SDCEGCGT", "EDUEDR04"
, "INCEGHH", "HWTEGBMI",
                  "HCUE_1AA", "SMKEDSTY", "ALCEDTYP", "CCCE_071", "CCCE_101", "PACEDPAI"
, "GEOEGPRV", "SDCEGRES", "DHHEGMS")
col_names_new <- c("heart", "osart", "age", "sex", "ethnicity", "education", "income",
"BMI",
                   "doctor", "smoker", "drinker", "highBP", "diabetes", "PAI", "provinc
e", "immigration", "marital")

osart11 <- cchs11 %>% filter(CCCA_05A == "OSTEOARTHRITIS" | CCCA_05A == "NOT APPLICABLE"
) %>% select(col_names_11)
osart21 <- cchs21 %>% filter(CCCC_05A == "OSTEOARTHRITIS" | CCCC_05A == "NOT APPLICABLE"
) %>% select(col_names_21)
osart31 <- cchs31 %>% filter(CCCE_05A == "OSTEOARTHRITIS" | CCCE_05A == "NOT APPLICABLE"
) %>% select(col_names_31)

names(osart11) <- col_names_new
names(osart21) <- col_names_new
names(osart31) <- col_names_new

osart_df <- do.call("rbind", list(osart11, osart21, osart31))
rm("osart11", "osart21", "osart31", "cchs11", "cchs21", "cchs31")

osart_df$osart <- revalue(osart_df$osart, c("OSTEOARTHRITIS"=1, "NOT APPLICABLE"=0))
osart_df$heart <- revalue(osart_df$heart, c("YES"=1, "NO"=0))
osart_df$doctor <- revalue(osart_df$doctor, c("YES"=1, "NO"=0))
osart_df$highBP <- revalue(osart_df$highBP, c("YES"=1, "NO"=0))
osart_df$diabetes <- revalue(osart_df$diabetes, c("YES"=1, "NO"=0))
osart_df$income <- revalue(osart_df$income, c("NO INCOME"="NO OR <$15,000", "LESS THAN 1
5,000"="NO OR <$15,000"))
osart_df$immigration <- revalue(osart_df$immigration, c("NOT APPLICABLE"="not immigrant"
, "0 TO 9 YEARS"="recent immigrant", "10 YEARS OR MORE"="more than 10 years", "10 OR MOR
E YEARS" = "more than 10 years"))
osart_df$education <- revalue(osart_df$education, c("OTHER POST-SEC."="POST-SEC.", "POST
-SEC. GRAD."="POST-SEC."))
osart_df$BMI <- cut(as.numeric(levels(osart_df$BMI))[osart_df$BMI], c(0, 18.5, 25, Inf),
                    labels=c('underweight', 'healthy', 'overweight'), right=FALSE)
osart_df$smoker <- mapvalues(osart_df$smoker, from = c("DAILY", "OCCASIONAL", "ALWAYS OC
CASION.", "FORMER DAILY", "FORMER OCCASION.", "NEVER SMOKED"), to = c("REGULAR", "OCCASI
ONAL", "OCCASIONAL", "FORMER", "FORMER", "NEVER"))
osart_df$marital <- revalue(osart_df$marital, c("SINGLE/NEVER MAR" = "SINGLE"))

osart_df <- replace(osart_df, osart_df=="NOT APPLICABLE", NA)
osart_df <- replace(osart_df, osart_df=="NOT STATED", NA)
```

```r
osart_df <- replace(osart_df, osart_df=="DON'T KNOW", NA)
osart_df <- replace(osart_df, osart_df=="REFUSAL", NA)
osart_df <- droplevels(osart_df)
osart_df <- na.omit(osart_df)

# summary of each column in the dataset
summary(osart_df)
```

```
##  heart      osart                 age            sex
##  1: 12016   1: 28162   40 TO 44 YEARS:27189   MALE  :115391
##  0:230804   0:214658   35 TO 39 YEARS:26655   FEMALE:127429
##                        30 TO 34 YEARS:24617
##                        45 TO 49 YEARS:23296
##                        50 TO 54 YEARS:23045
##                        25 TO 29 YEARS:21347
##                        (Other)       :96671
##             ethnicity                 education
##  WHITE           :219323   < THAN SECONDARY: 52829
##  VISIBLE MINORITY: 23497   SECONDARY GRAD. : 42120
##                            POST-SEC.       :147871
##
##
##
##
##              income               BMI         doctor
##  NO OR <$15,000 :22963   underweight:  7015   1:203317
##  $15,000-$29,999:40772   healthy    :112280   0: 39503
##  $30,000-$49,999:56039   overweight :123525
##  $50,000-$79,999:64261
##  $80,000 OR MORE:58785
##
##
##         smoker                  drinker       highBP     diabetes
##  REGULAR   : 53612   REGULAR DRINKER:153465   1: 36947   1: 11636
##  OCCASIONAL: 11708   OCC. DRINKER   : 44804   0:205873   0:231184
##  FORMER    :101243   FORMER DRINKER : 29622
##  NEVER     : 76257   NEVER DRANK    : 14929
##
##
##
##        PAI                     province                 immigration
##  ACTIVE  : 60432   ONTARIO         :75930   recent immigrant  :  6705
##  MODERATE: 62250   QUEBEC          :40520   more than 10 years: 24700
##  INACTIVE:120138   BRITISH COLUMBIA:29516   not immigrant     :211415
##                    ALBERTA         :23548
##                    MANITOBA        :13228
##                    SASKATCHEWAN    :13006
##                    (Other)         :47072
##           marital
##  MARRIED      :115138
##  COMMON-LAW   : 23894
##  WIDOW/SEP/DIV: 43588
##  SINGLE       : 60200
##
##
##
```

Q1

Within Canadian adults (20-64 years of age), is having osteoarthritis associated with the developing heart disease? For the purpose of this case study, assume that, from the literature, we know that the following variables are risk factors for the outcome and confounders in the above relationship: age, sex, ethnicity, education, household income, body mass index (BMI), access to a regular medical doctor, smoking habit, alcohol drinking habit, high-blood pressure, and diabetes. Also, assume that physical activity is suspected to be an intermediate factor between osteoarthritis and heart disease.

```
adult_range <- c("20 TO 24 YEARS", "25 TO 29 YEARS",  "30 TO 34 YEARS", "35 TO 39 YEARS"
,
  "40 TO 44 YEARS", "45 TO 49 YEARS", "50 TO 54 YEARS", "55 TO 59 YEARS", "60 TO 64 YEAR
S")

# take the average age for each age group
adult_numeric <- c(22, 27, 32, 37, 42, 47, 52, 57, 62)
osart_df <- osart_df[osart_df$age %in% adult_range, ]
osart_df$age <- as.numeric(mapvalues(osart_df$age, from = adult_range, to = adult_numeri
c))
```

```
# change base references
osart_df <- within(osart_df, heart <- relevel(heart, ref = "0"))
osart_df <- within(osart_df, osart <- relevel(osart, ref = "0"))
osart_df <- within(osart_df, BMI <- relevel(BMI, ref = "healthy"))
osart_df <- within(osart_df, smoker <- relevel(smoker, ref = "NEVER"))

lrmod <- glm(heart ~ osart + age + sex + ethnicity + education + income + BMI + doctor +
smoker + drinker + highBP + diabetes + PAI, data = osart_df, family = binomial)

summary(lrmod)
```

```
## 
## Call:
## glm(formula = heart ~ osart + age + sex + ethnicity + education + 
##     income + BMI + doctor + smoker + drinker + highBP + diabetes + 
##     PAI, family = binomial, data = osart_df)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -1.409  -0.243  -0.156  -0.107   3.725  
## 
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)    
## (Intercept)               -4.20529    0.11035  -38.11  < 2e-16 ***
## osart1                     0.39836    0.03518   11.32  < 2e-16 ***
## age                        0.31948    0.00777   41.12  < 2e-16 ***
## sexFEMALE                 -0.63509    0.02972  -21.37  < 2e-16 ***
## ethnicityVISIBLE MINORITY -0.16035    0.05470   -2.93   0.0034 ** 
## educationSECONDARY GRAD.  -0.14360    0.04421   -3.25   0.0012 ** 
## educationPOST-SEC.        -0.10845    0.03443   -3.15   0.0016 ** 
## income$15,000-$29,999     -0.33300    0.04698   -7.09  1.4e-12 ***
## income$30,000-$49,999     -0.56380    0.04573  -12.33  < 2e-16 ***
## income$50,000-$79,999     -0.69581    0.04713  -14.76  < 2e-16 ***
## income$80,000 OR MORE     -0.73403    0.05036  -14.58  < 2e-16 ***
## BMIunderweight             0.31413    0.11288    2.78   0.0054 ** 
## BMIoverweight              0.09536    0.03074    3.10   0.0019 ** 
## doctor0                   -0.63127    0.05286  -11.94  < 2e-16 ***
## smokerREGULAR              0.41278    0.04259    9.69  < 2e-16 ***
## smokerOCCASIONAL           0.37702    0.07605    4.96  7.1e-07 ***
## smokerFORMER               0.29553    0.03793    7.79  6.6e-15 ***
## drinkerOCC. DRINKER        0.24385    0.03691    6.61  3.9e-11 ***
## drinkerFORMER DRINKER      0.33947    0.03830    8.86  < 2e-16 ***
## drinkerNEVER DRANK         0.30891    0.06971    4.43  9.4e-06 ***
## highBP0                   -1.01939    0.03058  -33.34  < 2e-16 ***
## diabetes0                 -0.69019    0.04090  -16.87  < 2e-16 ***
## PAIMODERATE                0.04584    0.04146    1.11   0.2689    
## PAIINACTIVE                0.09174    0.03632    2.53   0.0115 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 53458  on 200001  degrees of freedom
## Residual deviance: 44607  on 199978  degrees of freedom
## AIC: 44655
## 
## Number of Fisher Scoring iterations: 7
```

From the logistic regression model, we can conclude that exposure to osteoarthritis patients are more likely to have cardiovascular diseases as well. Accounting for possible risk factors and confounders as mentioned above, osteoarthritis is a statistically significant contributor to the risk of heart diseases. Specifically, the model estimates that the odds of having cardiovascular disease will increase by a factor of \(exp(0.399) = 1.49\) when the patient has osteoarthritis. Note that an adjusted model is presented for prediction purposes in the last section, where the insignificant variable PAI is removed.

Q2

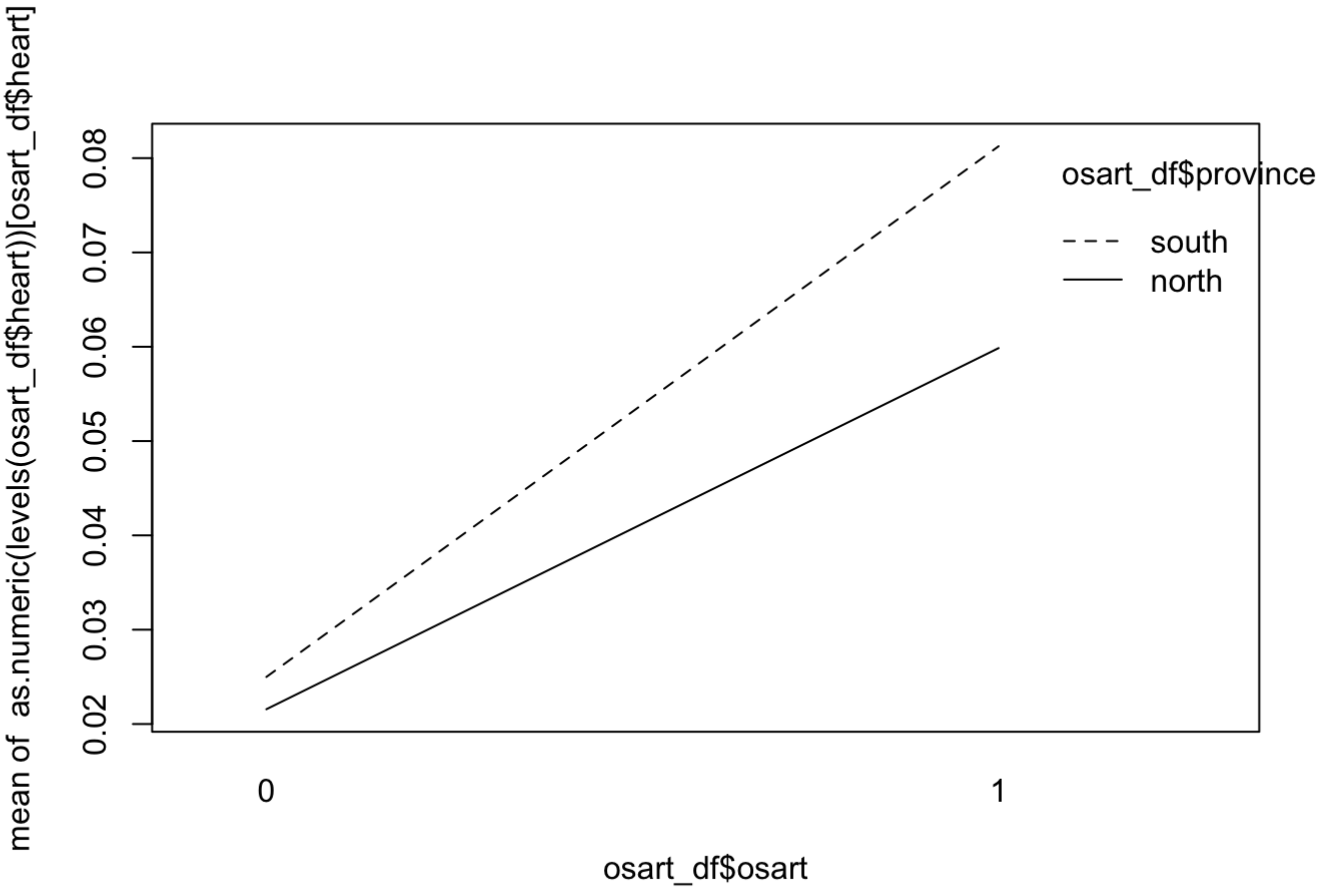
Does the relationship between osteoarthritis and heart disease vary: (a) between participants living in the northern parts of Canada versus those living in the southern parts?

```
osart_df$province <- mapvalues(osart_df$province,
          from = c("NEWFOUNDLAND", "PEI", "NOVA SCOTIA","NEW BRUNSWICK", "QU\xc9BEC",
                   "ONTARIO", "MANITOBA","SASKATCHEWAN", "ALBERTA", "BRITISH COLUMBIA",
                   "YUKON/NWT/NUNAVT", "NFLD & LAB.","QUEBEC", "YUKON/NWT/NUNA."),
          to = c(rep("south", 10), "north", "south", "south", "north"))
```

```
lr_prov <- glm(heart ~ osart + province + osart:province, data = osart_df, family = bino
mial)
tidy(lr_prov)
```

```
## # A tibble: 4 x 5
##   term                 estimate std.error statistic p.value
##   <chr>                   <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)            -3.66     0.0152  -242.      0
## 2 osart1                  1.24     0.0321    38.6     0
## 3 provincenorth          -0.151    0.106     -1.42    0.157
## 4 osart1:provincenorth   -0.178    0.273     -0.651   0.515
```

```
interaction.plot(x.factor = osart_df$osart,
                 trace.factor = osart_df$province,
                 response = as.numeric(levels(osart_df$heart))[osart_df$heart])
```



```
summary(aov(as.numeric(levels(osart_df$heart))[osart_df$heart] ~ osart*province, data =
osart_df))
```

```
##                   Df Sum Sq Mean Sq F value Pr(>F)
## osart          1e+00     49    48.9 1712.83 <2e-16 ***
## province       1e+00      0     0.1    3.23  0.072 .
## osart:province 1e+00      0     0.1    2.96  0.085 .
## Residuals      2e+05   5712     0.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-values for the location coefficient and the interaction coefficient are not significant. This means that where the patient is from does not affect the relationship between osteoarthritis and heart diseases. The plot also shows a parallel relationships between the lines with no interaction effect.

b. between men and women?

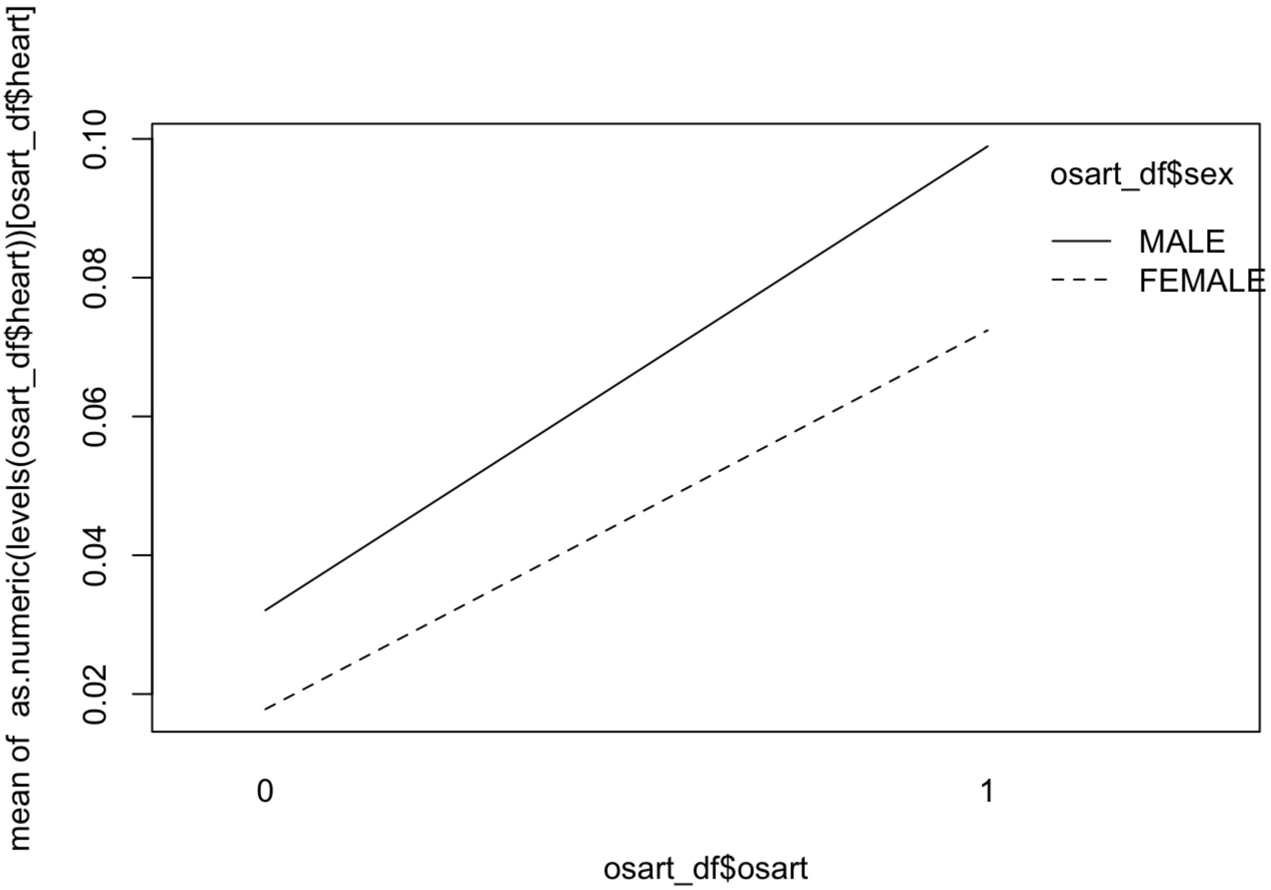
```
osart_df <- within(osart_df, sex <- relevel(sex, ref = "FEMALE"))
lr_sex <- glm(heart ~ osart + sex + osart:sex, data = osart_df, family = binomial)
tidy(lr_sex)
```

```
## # A tibble: 4 x 5
##   term           estimate std.error statistic   p.value
##   <chr>             <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)      -4.01     0.0249   -161.   0.
## 2 osart1            1.46     0.0436     33.4  3.17e-245
## 3 sexMALE           0.603    0.0312     19.3  5.12e- 83
## 4 osart1:sexMALE   -0.262    0.0657     -3.99 6.72e-  5
```

The logistic regression we consider here is: \[log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 * orart + \beta_2*sex + \beta_3 *orart*sex\] The p-values for all coefficients are statistically significant. This means that the association between heart diseases and osteoarthritis depends on the gender of the patient. Specifically, \(exp(\beta_0)/(1+exp(\beta_0)) = 0.018\) is the probability of having heart diseases when the patient is a female without osteoarthritis. \(exp(\beta_1) = exp(1.46) = 4.31\) is the odds ratio of heart disease comparing osteoarthritis and non-osteoarthritis among female patients. \(exp(\beta_2) = exp(0.603) = 1.83\) is the odds ratio of heart disease comparing males with females among non-osteoarthritis patients. \(exp(\beta_3) = exp(-0.26) = 0.77\) is the difference between the log-odds ratio comparing osteoarthritis vs. non-osteoarthritis in males and log-odds ratio comparing osteoarthritis vs. non-osteoarthritis in females. i.e.:

\[log\frac{odd_{o,m}}{odd_{no,m}} - log\frac{odd_{o,f}}{odd_{no,f}} = -0.26 = log\frac{odd_{o,m} * odd_{no,f}}{odd_{no,m}*odd_{o,f}}\] Therefore, \[exp(-0.26) = 0.77 = \frac{odd_{o,m}}{odd_{o,f}}/\frac{odd_{no,m}}{odd_{no,f}} = \frac{odd_{o,m}}{odd_{no,m}}/\frac{odd_{o,f}}{odd_{no,f}}\]

```
interaction.plot(x.factor = osart_df$osart,
                 trace.factor = osart_df$sex,
                 response = as.numeric(levels(osart_df$heart))[osart_df$heart])
```



```
summary(aov(as.numeric(levels(osart_df$heart))[osart_df$heart] ~ osart*sex, data = osart
_df))
```

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## osart       1e+00     49    48.9  1716.4 < 2e-16 ***
## sex         1e+00     11    11.4   400.0 < 2e-16 ***
## osart:sex   1e+00      1     0.5    18.1 2.1e-05 ***
## Residuals   2e+05   5700     0.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

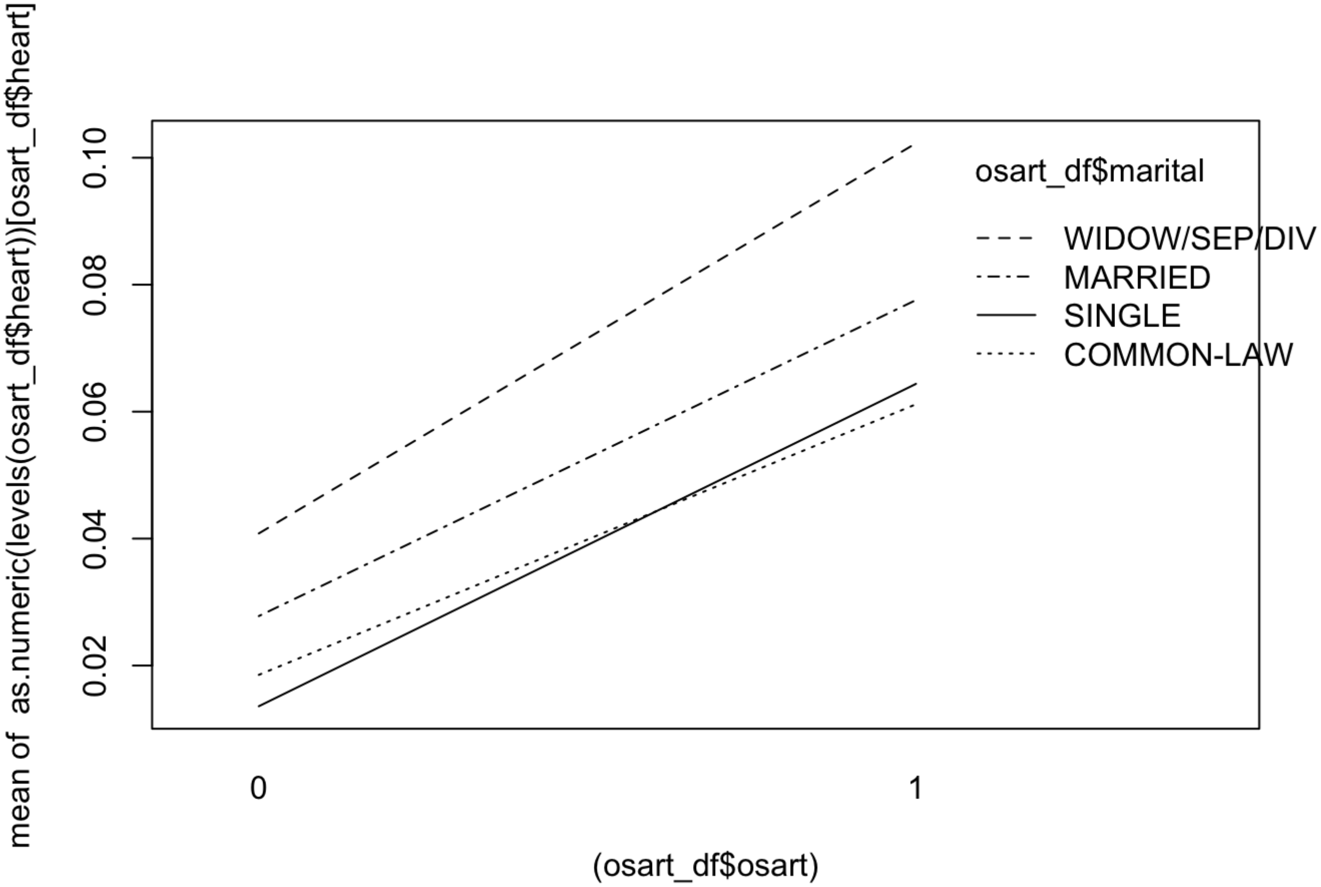
The results seem to suggest that male patients will have a lower chance of having heart diseases if they also have osteoarthritis. However, upon examining the parallel interaction plot, there is an absense of synergy between gender and osteoarthritis. It shows that statistical significance can occur based on the sheer size of the dataset. Nonetheless, the logistic regression model shows that male are more prone to having heart diseases.

c. by marital status?

```
lr_marital <- glm(heart ~ osart + marital + osart:marital, data = osart_df, family = bin
omial)
tidy(lr_marital)
```

```
## # A tibble: 8 x 5
##   term                         estimate std.error statistic   p.value
##   <chr>                           <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                   -3.55      0.0202   -176.   0.
## 2 osart1                         1.08      0.0437     24.7  1.16e-134
## 3 maritalCOMMON-LAW             -0.415     0.0539     -7.70 1.41e- 14
## 4 maritalWIDOW/SEP/DIV           0.398     0.0380     10.5  1.28e- 25
## 5 maritalSINGLE                 -0.730     0.0453    -16.1  1.79e- 58
## 6 osart1:maritalCOMMON-LAW       0.159     0.135       1.18 2.38e-  1
## 7 osart1:maritalWIDOW/SEP/DIV  -0.0941    0.0740     -1.27 2.04e-  1
## 8 osart1:maritalSINGLE           0.529     0.106       5.02 5.24e-  7
```

```
interaction.plot(x.factor = (osart_df$osart),
                 trace.factor = osart_df$marital,
                 response = as.numeric(levels(osart_df$heart))[osart_df$heart])
```



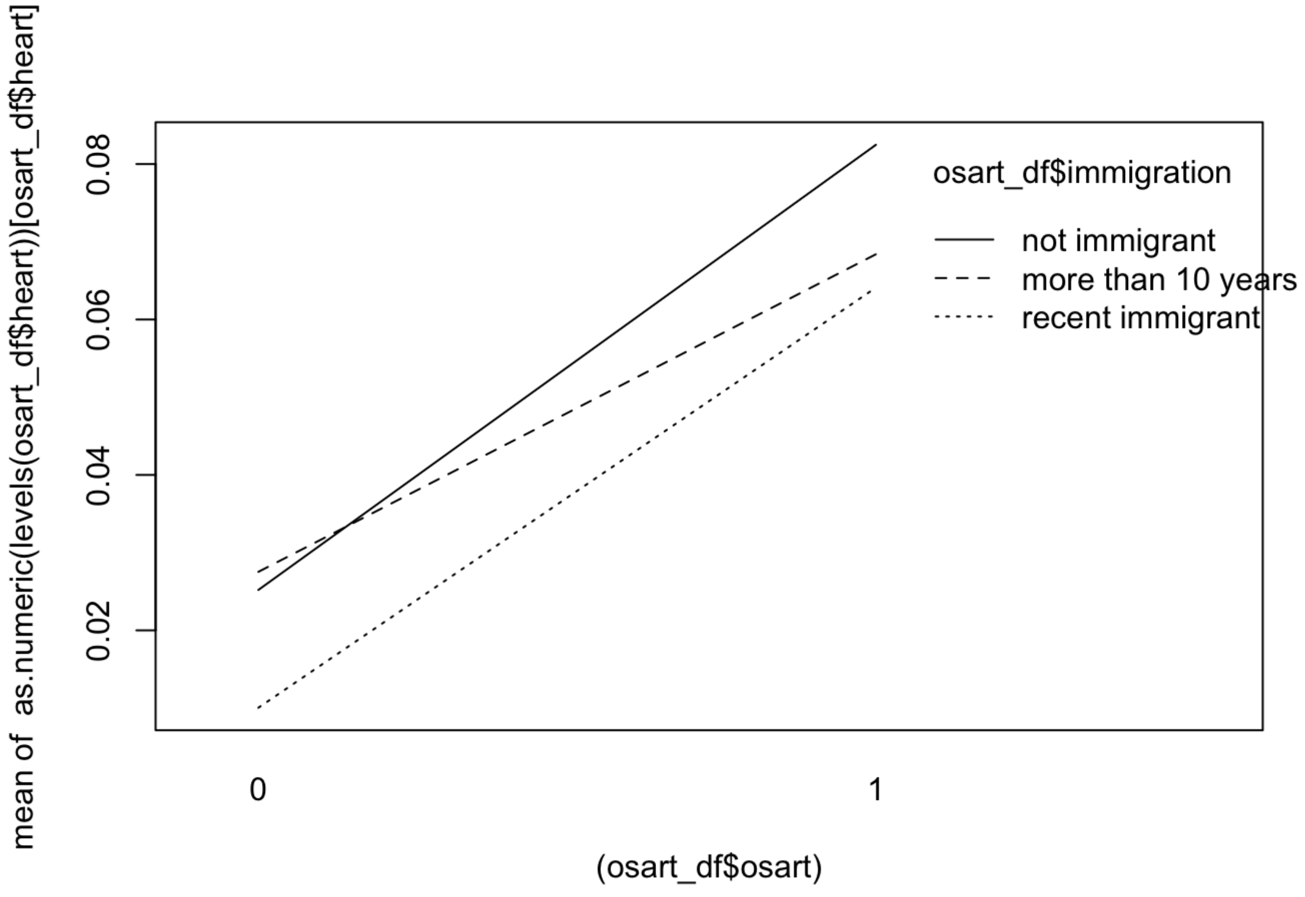
Given the same osteoarthritis status, the model suggests that compared to a married person, people with common-law status or single status have less chance of developing heart diseases, whereas people who are divided with their spouses have higher risks of heart diseases. The is only one interaction term that is significant. The interation plot suggests that we should not take this interaction seriously.

d. by recency of immigration?

```
lr_immigration <- glm(heart ~ osart + immigration + osart:immigration, data = osart_df,
family = binomial)
tidy(lr_immigration)
```

```
## # A tibble: 6 x 5
##   term                               estimate std.error statistic   p.value
##   <chr>                                 <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)                          -4.59      0.128    -36.0  3.03e-283
## 2 osart1                                1.91      0.480      3.98 6.84e-  5
## 3 immigrationmore than 10 years         1.03      0.136      7.55 4.35e- 14
## 4 immigrationnot immigrant              0.935     0.129      7.27 3.59e- 13
## 5 osart1:immigrationmore than 10 ye…   -0.957     0.491     -1.95 5.13e-  2
## 6 osart1:immigrationnot immigrant      -0.663     0.481     -1.38 1.68e-  1
```

```
interaction.plot(x.factor = (osart_df$osart),
                 trace.factor = osart_df$immigration,
                 response = as.numeric(levels(osart_df$heart))[osart_df$heart])
```



The immigration model shows that controlling for osteoarthritis exposure, older immigrants and non-immigrants both have a higher risk for heart diseases (by factors of 2.77 and 2.54 respectively) than recent immigrants. It is also statistically significant that among osteoarthritis patients, older immigrants are more likely to have heart diseases than recent immigrants. These results might be plausible because immigrants who have been to the country longer are probably older as well, thus more likely to catch diseases in general. On the other hand, non-immigrants might not have the same level of medical access as those who have immigration status. Again, the plot seems to suggest that intreaction terms are unnecessary since lines are relatively parallel.

Q4

With the information provided in the PUMF, what would be your interpretation of the analysis results? What are the limitations of this study? What additional information would be helpful in reaching a more meaningful conclusion?

Interpretations:

The analysis above clearly shows that osteoarthritis patients are more prone to heart diseases. On top of that, gender, age and marital status played significant role in the link between osteoarthritis and cardiovascular diseases. It is important to control for these variables if we do further experiments on the association between osteoarthritis and heart diseases.

Limitations:

The limitation of this study is that there are many possible combination of the interactions between potential contributors of heart diseases and confounders. To uncover every possible combination is somewhat infeasible with a limited amount of time. Some of the variables can be correlated as well, for instance, between marital status and age or immigration status and age. Another important issue is that we are only looking for correlation, not causation. A simple reversal of the osteoarthritis indicator and heart disease indicator in the model also yields statistically significant result. The above evidences cannot be used to conclude that osteoarthritis is a cause for cardiovasucular diseases. Lastly, we have only applied logistic regression to the dataset. There are other methods such as random forest or neural networks that might produce better results and provide new knowledge for us.

Additional Information Needed:

Less missing data can certainly be helpful to the analysis. It will also be helpful to control for other variables like age and gender in order to isolate the association between osteoarthritis and cardiovasucular diseases, although variabilities in physical and mental conditions among patients are hard to manage. Furthermore, the datasets are from 15 years ago, if we have access to the same patients now and see what have changed for them, it can provide insights for us as to what happens over time. For instance, given new data, it might be possible that heart diseases can be prevented if osteoarthritis is cured over time.

b)

Evaluate the predictive accuarcy of the model used to calculate the adjusted measure of association between osteoarthitis and heart disease. Do you recommend using this model to prdedict heart disease for Canadians?

Based on the above analysis, we remove the insignificant variables and purpose the following adjusted model below where all vairables are statistically significant and no interaction terms are present:

```
lr_adjusted <- glm(heart ~ osart + age + sex + ethnicity + education + income + BMI + do
ctor + smoker + drinker + highBP + diabetes, data = osart_df, family = binomial)

summary(lr_adjusted)
```

```
## 
## Call:
## glm(formula = heart ~ osart + age + sex + ethnicity + education + 
##     income + BMI + doctor + smoker + drinker + highBP + diabetes, 
##     family = binomial, data = osart_df)
## 
## Deviance Residuals: 
##    Min      1Q  Median      3Q     Max  
## -1.400  -0.243  -0.156  -0.107   3.735  
## 
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                 -4.77606    0.10631  -44.92  < 2e-16 ***
## osart1                       0.39988    0.03517   11.37  < 2e-16 ***
## age                          0.31942    0.00776   41.15  < 2e-16 ***
## sexMALE                      0.63315    0.02970   21.32  < 2e-16 ***
## ethnicityVISIBLE MINORITY -0.15777    0.05469   -2.88  0.00391 ** 
## educationSECONDARY GRAD.  -0.14714    0.04418   -3.33  0.00087 ***
## educationPOST-SEC.        -0.11482    0.03435   -3.34  0.00083 ***
## income$15,000-$29,999     -0.33174    0.04697   -7.06  1.6e-12 ***
## income$30,000-$49,999     -0.56309    0.04573  -12.31  < 2e-16 ***
## income$50,000-$79,999     -0.69698    0.04713  -14.79  < 2e-16 ***
## income$80,000 OR MORE     -0.73831    0.05033  -14.67  < 2e-16 ***
## BMIunderweight             0.31714    0.11288    2.81  0.00496 ** 
## BMIoverweight              0.10037    0.03068    3.27  0.00107 ** 
## doctor0                   -0.62947    0.05285  -11.91  < 2e-16 ***
## smokerREGULAR              0.42085    0.04248    9.91  < 2e-16 ***
## smokerOCCASIONAL           0.37840    0.07604    4.98  6.5e-07 ***
## smokerFORMER               0.29572    0.03792    7.80  6.3e-15 ***
## drinkerOCC. DRINKER        0.24810    0.03687    6.73  1.7e-11 ***
## drinkerFORMER DRINKER      0.34310    0.03825    8.97  < 2e-16 ***
## drinkerNEVER DRANK         0.31537    0.06966    4.53  6.0e-06 ***
## highBP0                   -1.02236    0.03056  -33.46  < 2e-16 ***
## diabetes0                 -0.69299    0.04089  -16.95  < 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: 53458  on 200001  degrees of freedom
## Residual deviance: 44614  on 199980  degrees of freedom
## AIC: 44658
## 
## Number of Fisher Scoring iterations: 7
```

```
pred_probs <- lr_adjusted %>% predict(type = "response")
# 0.111 achieves the highest f_score
pred_class  <- ifelse(pred_probs >= 0.1, "positive", "negative")
confusion <- table(osart_df$heart, pred_class)
tp <- confusion[2,2]
fp <- confusion[1,2]
fn <- confusion[2,1]
tn <- confusion[1,1]
recall <- tp / (fn + tp)
precision <- tp /(tp + fp)
accuracy <- (tp + tn) / (tp + tn + fp + fn)
f_score <- (2*precision*recall)/(recall+precision)
confusion
```

```
##    pred_class
##     negative positive
##   0   183727    10338
##   1     3745     2192
```

```
sprintf("accuracy: %.4f", accuracy)
```

```
## [1] "accuracy: 0.9296"
```

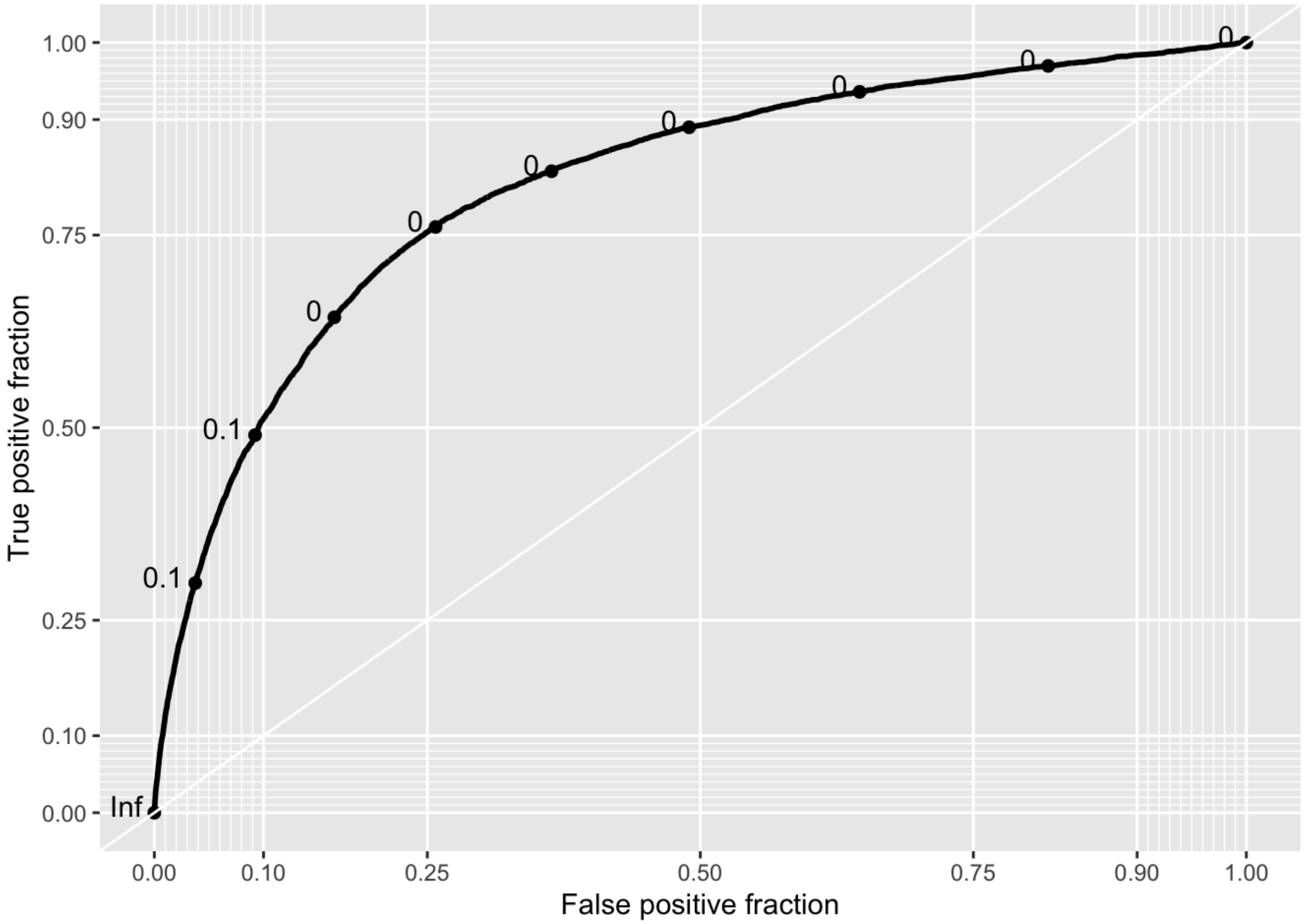
```
sprintf("F-score: %.4f", f_score)
```

```
## [1] "F-score: 0.2374"
```

```
sprintf("false negative probability: %.4f", fn/(tp + tn + fp + fn))
```

```
## [1] "false negative probability: 0.0187"
```

```
tibble(pred = pred_probs, obs = as.numeric(levels(osart_df$heart))[osart_df$heart]) %>%
  ggplot(aes(d = obs, m = pred)) + geom_roc() + style_roc(theme = theme_gray)
```



In the adjusted logistic regression model, we exclude the physical activity index (PAI) compared to the full model because PAI was not statistically significant. The adjusted model has a decent-looking concave ROC curve. Even though the model can achieve a good accuracy overall, it seems to have overfitted the data. This is because when we increase the threshold towards 1, we see that the model is essentially guessing towards every patient not having heart diseases. This is because over 95% of the patients in the dataset do not have a heart problem, and just by guessing that the patient is free of cardiovascular diseases all the time can achieve a high accuracy. Therefore, we should look into more detailed metrics like the false negative rate. We care about false negative probability because it is more important to reduce the number of people who are predicted to have no heart problem but in fact do (false negatives) than to wrongly predict someone healthy to have heart diseases (false positives). When the threshold is set to \(0.1\), the model achieves an accuracy of \(93\%\) and a F-score of \(0.24\) while maintaining a low false negative probability around \(1.9\%\). I recommend using this logistic regression model with the threshold of \(0.1\) to predict heart diseases in similar patients. However, we need to be cautious about the false negative probability and keep in mind the bias present in the dataset.