A hospital audit(P#2)

Introduction

This is the result of auditting the hospital's breast cancer diagnostic screening. There are 5 radiologists, and they conducted 987 screenings by reading mammograms.

```
library(tidyverse)
library(mosaic)
library(knitr)
library(kableExtra)
```

```
# data
brca = read.csv("../data/brca.csv")
head(brca)
```

```
##     radiologist cancer recall       age history symptoms    menopause
## 1 radiologist13      0      1   age4049       0        0      premeno
## 2 radiologist13      0      0   age6069       1        0 postmenoNoHT
## 3 radiologist13      0      0   age5059       0        0 postmenoNoHT
## 4 radiologist13      0      0   age5059       0        0   postmenoHT
## 5 radiologist13      0      0 age70plus       0        0   postmenoHT
## 6 radiologist13      0      0   age6069       0        0   postmenoHT
##    density
## 1 density3
## 2 density3
## 3 density3
## 4 density3
## 5 density1
## 6 density3
```

First, I review the 5 radiologists' recalling history, and see who is more conservative to recall the patients. Second, I compare the radiologists' interpretations of mammpgrams with risk-factor-combined results.

Conservative Rate : who is more conservative to recall the patient

Simple result

We can simply think that those who recall the patient frequently are more conservative. Following is the result by the idea. The radiologist89 is the most conservative radiologit.

```
xtabs(~recall + radiologist, data = brca) %>%
  prop.table(margin = 2) %>% round(3) %>% kable() %>% kable_styling("striped", full_width = TRUE)
```

	radiologist13	radiologist34	radiologist66	radiologist89	radiologist95
0	0.854	0.914	0.813	0.807	0.863
1	0.146	0.086	0.187	0.193	0.137

However, the patients radiologists examined are different. So, some are more likely to develop cancer, which makes the radiologist recalls the patient more frequently. Thus, I will make a model to predict cancer with risk factors, then deweight the recalling rate by the cancer probability of the patient, and calculate the conservative rate.

Cancer model

Preparation

Before I build the model, check the risk factors with cancer probability.

- The relationship between age and cancer

```
xtabs(~cancer + age, data = brca) %>% prop.table(margin=2) %>%
  round(3) %>% kable() %>% kable_styling("striped", full_width = TRUE)
```

	age4049	age5059	age6069	age70plus
0	0.969	0.965	0.975	0.94
1	0.031	0.035	0.025	0.06

- The relationship between history and cancer

```
xtabs(~cancer + history, data = brca) %>% prop.table(margin=2) %>%
  round(3) %>% kable() %>% kable_styling("striped", full_width = TRUE)
```

	0	1
0	0.964	0.954
1	0.036	0.046

- The relationship between symptoms and cancer

```
xtabs(~cancer + symptoms, data = brca) %>% prop.table(margin=2) %>%
  round(3) %>% kable() %>% kable_styling("striped", full_width = TRUE)
```

	0	1
0	0.964	0.938
1	0.036	0.062

- The relationship between menopause and cancer

```
xtabs(~cancer + menopause, data = brca) %>% prop.table(margin=2) %>%
  round(3) %>% kable() %>% kable_styling("striped", full_width = TRUE)
```

	postmenoHT	postmenoNoHT	postmenounknown	premeno
0	0.963	0.967	0.914	0.963
1	0.037	0.033	0.086	0.037

- The relationship density symptoms and cancer

```
xtabs(~cancer + density, data = brca) %>% prop.table(margin=2) %>%
  round(3) %>% kable() %>% kable_styling("striped", full_width = TRUE)
```

	density1	density2	density3	density4
0	0.989	0.967	0.963	0.925
1	0.011	0.033	0.037	0.075

We can see that those who over 70 years old, or postmenounknown, or density 4 are likely to develop cancer. I make dummy variables for those features.

```
brca = mutate(brca, old = ifelse(brca$age=='age70plus', 1, 0),
              menop = ifelse(brca$menopause=='postmenounknown', 1, 0),
              dense = ifelse(brca$density=='density4', 1, 0))
```

Last, I define a functions to evaluate the model. I will use deviance for evaluation.

```
dev_out = function(y, probhat) {
  rc_pairs = cbind(seq_along(y), y)
  -2*sum(log(probhat[rc_pairs]))
}
```

modeling

I build a model with train data(80%), and test the model with test data(20%). The train data are randomly selected, so I repeat 100 times the random sampling and modeling, and average the test result. I try 5 models at the same time. For a benchmark, I use the average cancer probability. With many trials, I make the models which are better than the benchmark. Among them, Model 3 has the least deviance.

```
### split 80% train and 20% test
n = nrow(brca)
n_train = round(0.8*n)
n_test = n - n_train

### averaging test result
err_grid = do(100) * {
  #### split train and test set
  train_cases = sample.int(n, n_train, replace = FALSE)
  test_cases = setdiff(1:n, train_cases)
  brca_train = brca[train_cases,]
  brca_test = brca[test_cases,]

  #### bench mark; using a global cancer probability
  p_bm = length(which(brca_test$cancer=="1"))/n_test
  bm = rep(p_bm, times = n_test)

  #### logit
  logit1 = glm(cancer ~ . - recall - radiologist - old - menop - dense, data = brca_train, family = 'bi
  logit2 = glm(cancer ~ history + symptoms + old + menop + dense, data = brca_train, family = 'binomial
  logit3 = glm(cancer ~ history, data = brca_train, family = 'binomial')
  logit4 = glm(cancer ~ history + symptoms, data = brca_train, family = 'binomial')
  logit5 = glm(cancer ~ history + symptoms + old, data = brca_train, family = 'binomial')

  ##### likelyhood
  phat1 = predict(logit1, brca_test, type = "response")
  phat2 = predict(logit2, brca_test, type = "response")
  phat3 = predict(logit3, brca_test, type = "response")
  phat4 = predict(logit4, brca_test, type = "response")
  phat5 = predict(logit5, brca_test, type = "response")

  #### deviance
  dev_bm = dev_out(brca_test$cancer, bm)
  dev_logit1 = dev_out(brca_test$cancer, phat1)
  dev_logit2 = dev_out(brca_test$cancer, phat2)
```

```
  dev_logit3 = dev_out(brca_test$cancer, phat3)
  dev_logit4 = dev_out(brca_test$cancer, phat4)
  dev_logit5 = dev_out(brca_test$cancer, phat5)

  c(dev_bm, dev_logit1, dev_logit2, dev_logit3, dev_logit4, dev_logit5)
}
dev_table = data.frame(
  Models = c("Cancer Prob(BM)", "Model1", "Model2", "Model3", "Model4", "Model5"),
  Deviances = colMeans(err_grid)
)
kable(t(dev_table)) %>%
  kable_styling("striped", full_width = TRUE) # select 'logit3' model
```

	V1	V2	V3	V4	V5	V6
Models	Cancer Prob(BM)	Model1	Model2	Model3	Model4	Model5
Deviances	1355.846	1559.871	1411.903	1350.273	1358.370	1378.767

Deweighting conservative rate

To compare the consevative rate, split the data for each radiologist.

```
brca_13 = brca[brca$radiologist=="radiologist13",]
brca_34 = brca[brca$radiologist=="radiologist34",]
brca_66 = brca[brca$radiologist=="radiologist66",]
brca_89 = brca[brca$radiologist=="radiologist89",]
brca_95 = brca[brca$radiologist=="radiologist95",]
```

I will deweight the consevative rate with the cancer probabilty by the model 3.

```
logit_test = glm(cancer ~ history, data = brca, family = 'binomial')
```

Define a function to calculate the deweighting conservative rate.

```
conserv = function(y) {
  p_canc = predict(logit_test, y, type = "response")
  recall_p_canc = y$recall/p_canc
  sum(recall_p_canc)/nrow(y)
} # more conservative for less cancer probabilities
```

Then, the conservative rates considering cancer probabiilty are as following. The result is the same as the simple result, the most conservative radiologist is radiologist89, and the others also the same.

```
conserv_table = data.frame(
  radiologist=c("radiologist13","radiologist34","radiologist66","radiologist89","radiologist95"),
  conservative_rate=c(conserv(brca_13), conserv(brca_34), conserv(brca_66), conserv(brca_89), conserv(br

kable(t(conserv_table)) %>%
  kable_styling("striped", full_width = TRUE) # radiologist89 is the most conservative
```

radiologist	radiologist13	radiologist34	radiologist66	radiologist89	radiologist95
conservative_rate	3.947362	2.387318	5.111808	5.152459	3.491379

Adding risk factors to interpret the mammogram

I compare the recall based cancer probability with the risk-factor based cancer probability. I split the data for train and test by random sampling. Then, repeat 100 times to bulid and test models, and average the results. The radiologists should be minimizing false negatives and also minimizing false positive. So, I use a True Positive Rate(TPR) and a False Positive Rate(FPR) to evaluate the models. High TPR means minimizing false negatives, low FPR means minimizing false positive. These are the functions I will use to test the models.

```
### function for TPR
TPR = function(y, yhat) {
  length(which(y==1 & yhat==1))/length(which(y==1))
  }

### function for FPR
FPR = function(y, yhat) {
  length(which(y==0 & yhat==1))/length(which(y==0))
}
```

This is the function to make yhat from phat.

```
### function for yhat
yhat = function(model, test_set){
  phat = predict(model, test_set, type = "response")
  ifelse(phat > 0.5, 1, 0)
}
```

There are 4 results, the first is recall-only result, the second is the result of recall and all factors model, the third is the result of recall and the best predicting model, the last is the result of the best predicting model. However, all predicting model can not be used to determine whether the patients should visit the hospital or not. The phat results are mostly zero. We can see that the radiolosists' interetation is the best for determining the recall of the patient.

```
test_grid = do(100) * {

  #### split train and test set
  train_cases = sample.int(n, n_train, replace = FALSE)
  test_cases = setdiff(1:n, train_cases)
  brca_train = brca[train_cases,]
  brca_test = brca[test_cases,]

  #### bench mark for confusion rate : cancer probabilty
  #p_bm = length(which(brca_test$cancer=="1"))/n_test
  #bm = rep(p_bm, times = n_test)

  #### logit
  M_rf1 = glm(cancer ~ .- radiologist - old - menop - dense, data = brca_train, family = 'binomial')
  M_rf2 = glm(cancer ~ recall + history, data = brca_train, family = 'binomial')
  M_rf3 = glm(cancer ~ history, data = brca_train, family = 'binomial')

  ##### likelyhood and
  yh_rf1 = yhat(M_rf1, brca_test)
  yh_rf2 = yhat(M_rf2, brca_test)
```

```
  yh_rf3 = yhat(M_rf3, brca_test)

  #### TPR, FPR
  c(TPR(brca_test$cancer,brca_test$recall), FPR(brca_test$cancer,brca_test$recall),
    TPR(brca_test$cancer,yh_rf1), FPR(brca_test$cancer,yh_rf1),
    TPR(brca_test$cancer,yh_rf2), FPR(brca_test$cancer,yh_rf2),
    TPR(brca_test$cancer,yh_rf3), FPR(brca_test$cancer,yh_rf3))
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
errMeans = colMeans(test_grid) %>% round(3)
err = matrix(errMeans, nrow=2, dimnames =
               list(c("TPR", "FPR"),
                    c("Recall", "Model1", "Model2", "Model3")))
kable(err) %>% kable_styling("striped", full_width = TRUE)
```

	Recall	Model1	Model2	Model3
TPR	0.564	0.007	0	0
FPR	0.133	0.001	0	0