131 hw3

Tianyi Li

2022-04-16

```r
library(ggplot2)
library(tidymodels)
library(ISLR)
library(ISLR2)
library(discrim)
library(poissonreg)
library(corrr)
library(klaR)
tidymodels_prefer()
```

```r
titanic <- read.csv("titanic.csv")
head(titanic)
```

```
##   passenger_id survived pclass
## 1            1       No      3
## 2            2      Yes      1
## 3            3      Yes      3
## 4            4      Yes      1
## 5            5       No      3
## 6            6       No      3
##                                                  name    sex age sib_sp parch
## 1                             Braund, Mr. Owen Harris   male  22      1     0
## 2 Cumings, Mrs. John Bradley (Florence Briggs Thayer) female  38      1     0
## 3                              Heikkinen, Miss. Laina female  26      0     0
## 4        Futrelle, Mrs. Jacques Heath (Lily May Peel) female  35      1     0
## 5                            Allen, Mr. William Henry   male  35      0     0
## 6                                    Moran, Mr. James   male  NA      0     0
##             ticket    fare cabin embarked
## 1        A/5 21171  7.2500  <NA>        S
## 2         PC 17599 71.2833   C85        C
## 3 STON/O2. 3101282  7.9250  <NA>        S
## 4           113803 53.1000  C123        S
## 5           373450  8.0500  <NA>        S
## 6           330877  8.4583  <NA>        Q
```

```r
titanic$survived<-factor(titanic$survived, levels=c('Yes','No'))
titanic$pclass<-factor(titanic$pclass)
```

Question 1

This dataset can be divided into different subgroups, and stratified sampling can generate representations more accurately of the population.

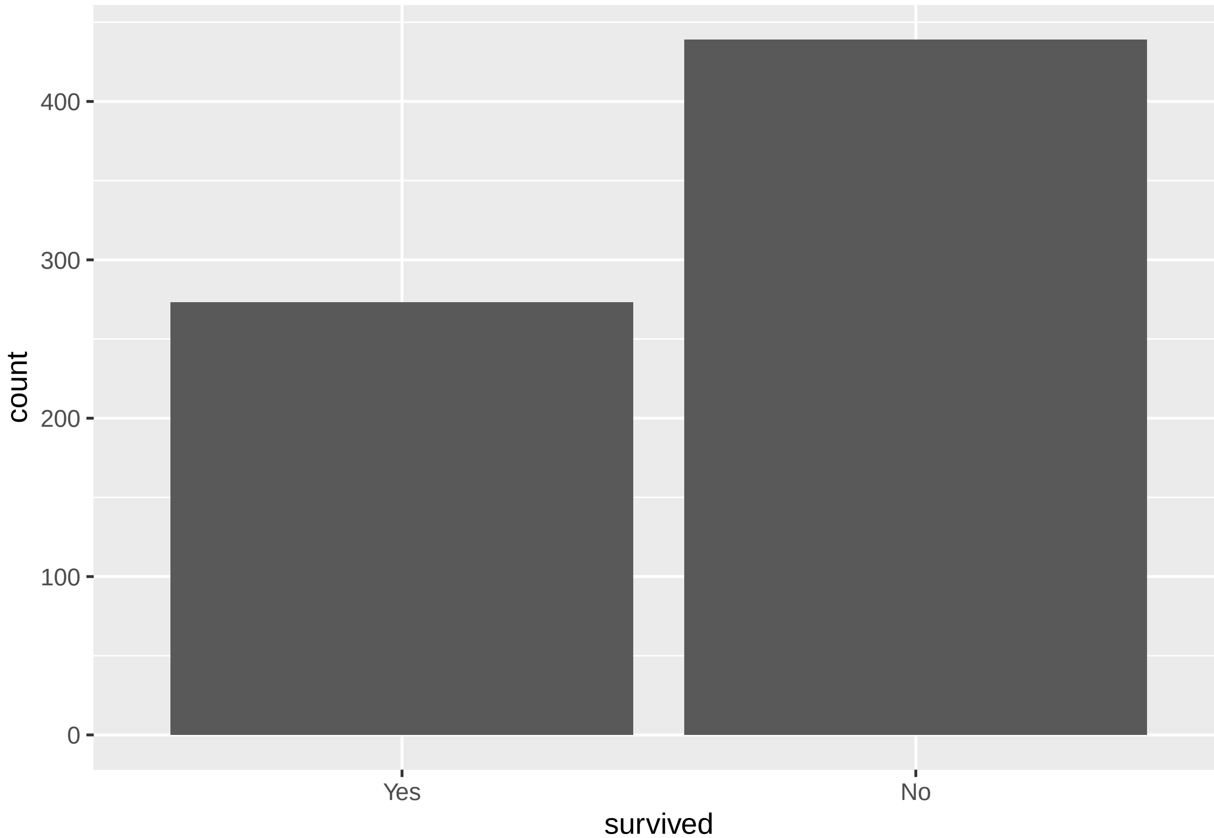
```r
set.seed(1202)
titanic_split<-initial_split(titanic,prop=0.80,
```

```
                              strata = survived )
titanic_train<-training(titanic_split)
titanic_test<-testing(titanic_split)
```

Question 2

The number of passengers who survived is significantly more than that of didn't survived.

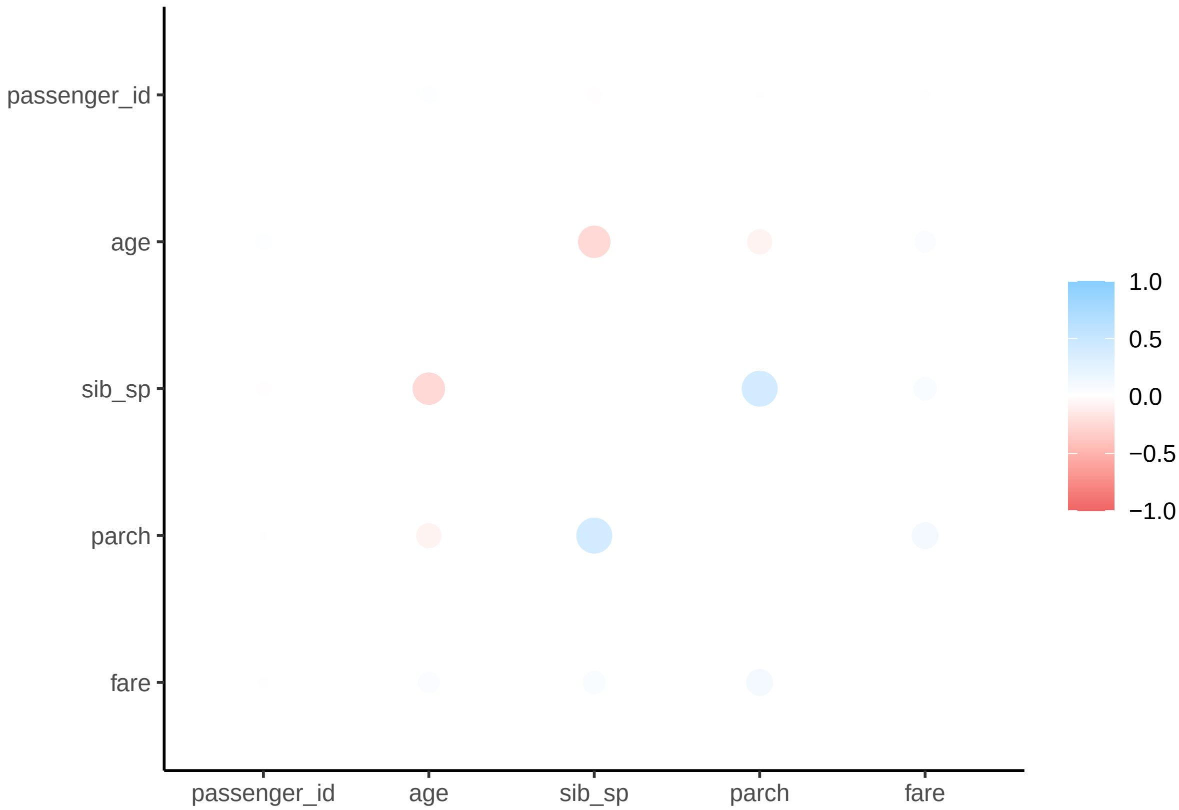
```
ggplot(titanic_train,aes(x=survived))+geom_bar()
```



Question 3

I see a symmetric and evenly distributed pattern. age and sib_sp are negatively correlated, parch and sib_sp are positively correlated.

```
cor_titanic <- titanic_train %>%
  select(is.numeric) %>%
  correlate()
rplot(cor_titanic)
```



Question 4

```
titanic_recipe <- recipe(survived ~ pclass + sex + age + sib_sp +
                           parch + fare, data = titanic_train) %>%
  step_impute_linear(age) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms= ~ starts_with("sex"):fare+
                  age:fare)
titanic_recipe
```

```
## Recipe
##
## Inputs:
##
##       role #variables
##    outcome          1
##  predictor          6
##
## Operations:
##
## Linear regression imputation for age
## Dummy variables from all_nominal_predictors()
## Interactions with starts_with("sex"):fare + age:fare
```

Question 5

```
log_reg <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")
log_wkflow <- workflow() %>%
```

```r
  add_model(log_reg) %>%
  add_recipe(titanic_recipe)
log_fit <- fit(log_wkflow, titanic_train)
```

Question 6

```r
lda_mod <- discrim_linear() %>%
  set_mode("classification") %>%
  set_engine("MASS")
lda_wkflow <- workflow() %>%
  add_model(lda_mod) %>%
  add_recipe(titanic_recipe)
lda_fit <- fit(lda_wkflow, titanic_train)
```

Question 7

```r
qda_mod <- discrim_quad() %>%
  set_mode("classification") %>%
  set_engine("MASS")
qda_wkflow <- workflow() %>%
  add_model(qda_mod) %>%
  add_recipe(titanic_recipe)
qda_fit <- fit(qda_wkflow, titanic_train)
```

Question 8

```r
nb_mod <- naive_Bayes() %>%
  set_mode("classification") %>%
  set_engine("klaR") %>%
  set_args(usekernel = FALSE)
nb_wkflow <- workflow() %>%
  add_model(nb_mod) %>%
  add_recipe(titanic_recipe)
nb_fit <- fit(nb_wkflow, titanic_train)
```

Question 9

Logistic Regression model achieved the highest accuracy.

```r
bind_titanic_train=bind_cols(predict(log_fit,new_data=titanic_train,type="class"),
                             predict(lda_fit,new_data=titanic_train,type="class"),
                             predict(qda_fit,new_data=titanic_train,type="class"),
                             predict(nb_fit,new_data=titanic_train,type="class"),
                             titanic_train$survived)
```

```r
log_reg_acc <- augment(log_fit, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
log_reg_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.810
```

```r
lda_acc <- augment(lda_fit, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
lda_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.801
```

```r
qda_acc <- augment(qda_fit, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
qda_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.781
```

```r
nb_acc <- augment(nb_fit, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)
nb_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.770
```

```r
accuracies <- c(log_reg_acc$.estimate, lda_acc$.estimate,
                nb_acc$.estimate, qda_acc$.estimate)
models <- c("Logistic Regression", "LDA", "Naive Bayes", "QDA")
results <- tibble(accuracies = accuracies, models = models)
results %>%
  arrange(-accuracies)
```

```
## # A tibble: 4 x 2
##   accuracies models
##        <dbl> <chr>
## 1      0.810 Logistic Regression
## 2      0.801 LDA
## 3      0.781 QDA
## 4      0.770 Naive Bayes
```

Question 10

The model performs fairly, not very accurately. The value differ because the model is optimized for the training accuracy.

```r
predict(log_fit, new_data = titanic_test, type = "prob")
```

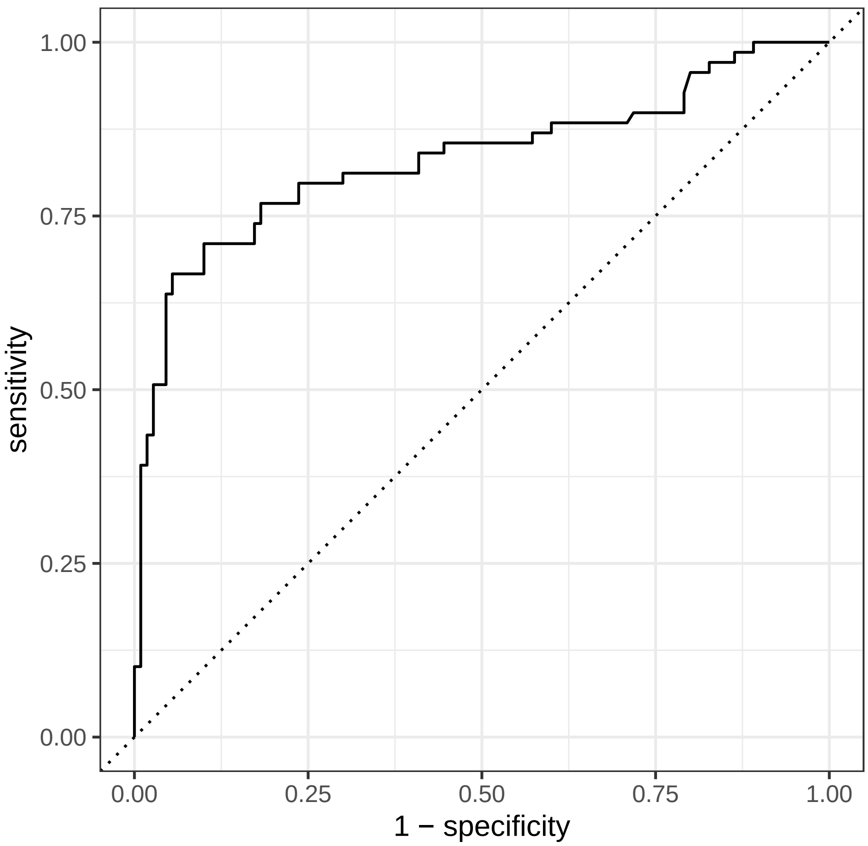
```
## # A tibble: 179 x 2
##    .pred_Yes .pred_No
##        <dbl>    <dbl>
##  1    0.940    0.0600
##  2    0.898    0.102
##  3    0.148    0.852
##  4    0.459    0.541
##  5    0.619    0.381
##  6    0.325    0.675
##  7    0.0946   0.905
##  8    0.979    0.0209
##  9    0.0484   0.952
## 10    0.315    0.685
```

```
## # ... with 169 more rows
```

```r
augment(log_fit, new_data = titanic_test) %>%
  conf_mat(truth = survived, estimate = .pred_class)
```

```
##           Truth
## Prediction Yes No
##        Yes  48 11
##        No   21 99
```

```r
augment(log_fit, new_data = titanic_test) %>%
  roc_curve(survived, .pred_Yes) %>%
  autoplot()
```




```r
augment(log_fit, new_data = titanic_test) %>%
  roc_auc(survived,.pred_Yes)
```

```
## # A tibble: 1 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.832
```