stat151_hw6

1.Kaggle competition

```
#data observation
train=read.csv('train.csv')
test=read.csv('test.csv')
ntrain=dim(train)[1]
ntrain
```

```
## [1] 891
```

```
ntest=dim(test)[1]
ntest
```

```
## [1] 418
```

```
full=rbind(train[,-2],test)
nfull=dim(full)[1]
nfull
```

```
## [1] 1309
```

```
#print info of all columns
ncol=dim(full)[2]
for(i in 1:ncol){
  cur=full[,i]
  message(colnames(full)[i],": ", class(cur))
  if(class(cur)=='factor'){
    message("    numbe of levels: ",length(levels(cur)))
  }
}
```

```
## PassengerId: integer

## Pclass: integer

## Name: factor

##     numbe of levels: 1307

## Sex: factor

##     numbe of levels: 2

## Age: numeric

## SibSp: integer

## Parch: integer

## Ticket: factor

##     numbe of levels: 929

## Fare: numeric

## Cabin: factor

##     numbe of levels: 187
```

```
## Embarked: factor
##     numbe of levels: 4
```

```r
sum(rowSums(is.na(full))>0)
```

```
## [1] 264
```

```r
sum(is.na(full$Age))
```

```
## [1] 263
```

```r
which(is.na(full$Fare))
```

```
## [1] 1044
```

```r
table(full$SibSp)
```

```
##
##   0   1   2   3   4   5   8
## 891 319  42  20  22   6   9
```

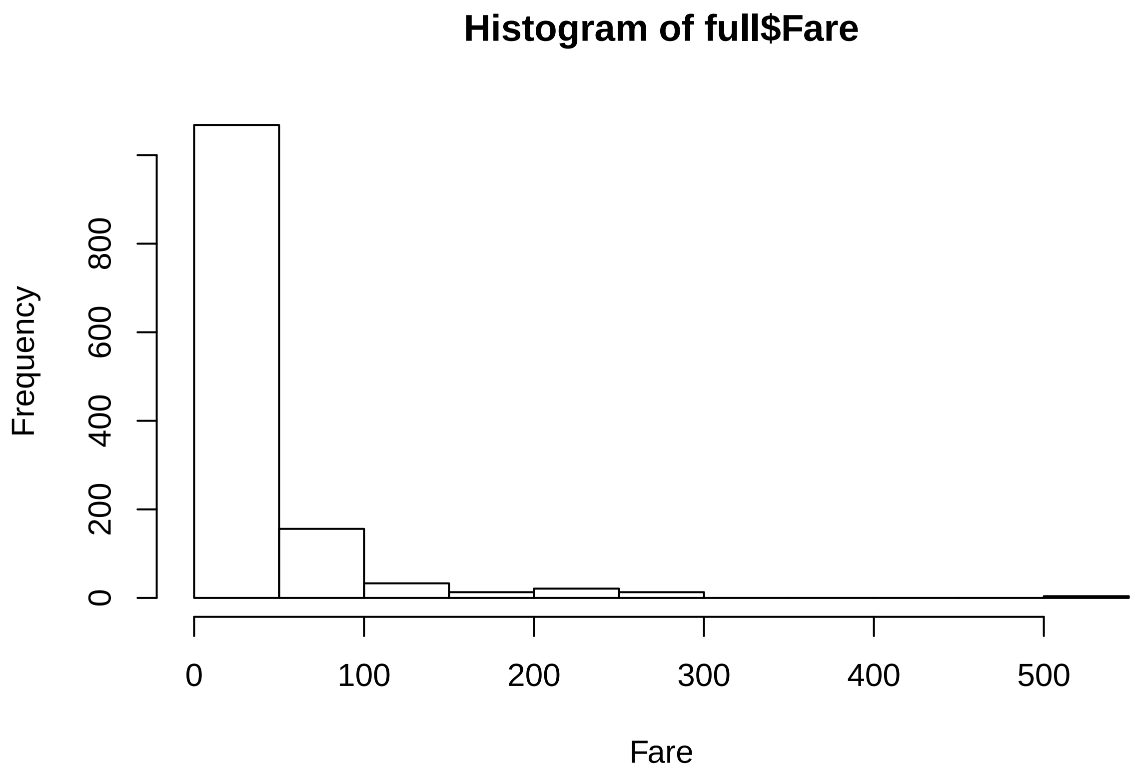
```r
table(full$Parch)
```

```
##
##    0    1    2    3    4    5    6    9
## 1002  170  113    8    6    6    2    2
```

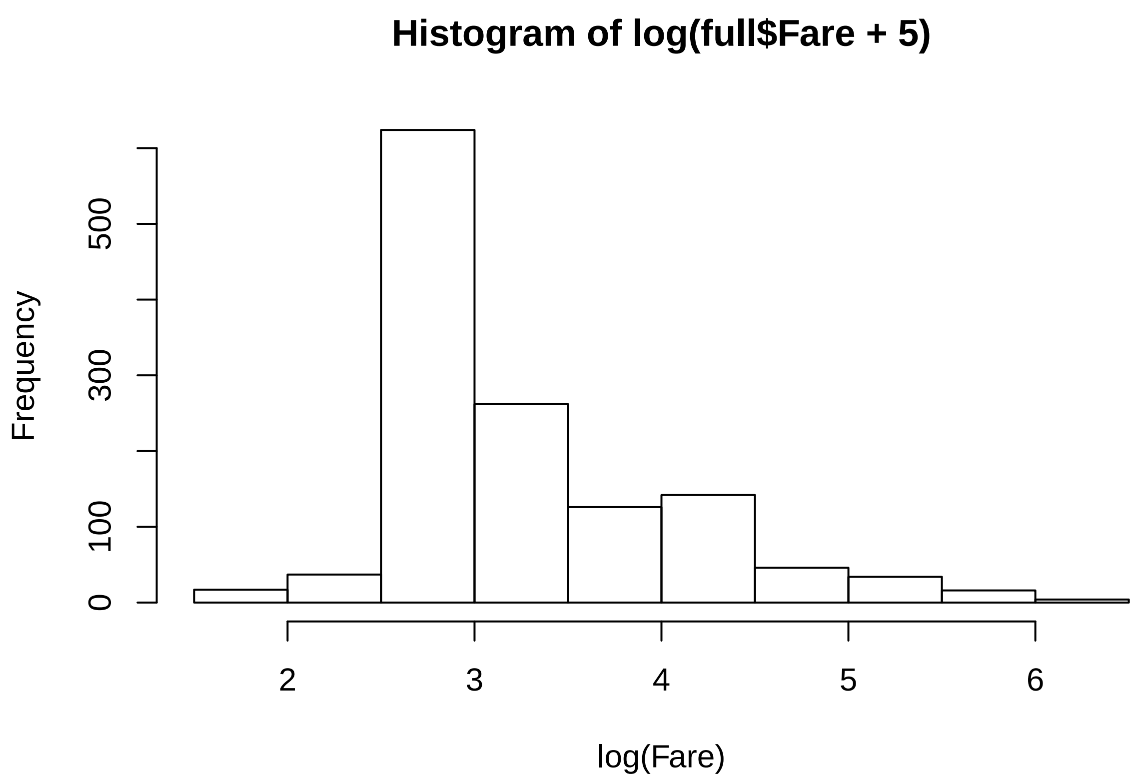
```r
table(table(full$Ticket))
```

```
##
##   1   2   3   4   5   6   7   8  11
## 713 132  49  16   7   4   5   2   1
```

```r
 #according to the observation, we would convert Pclass into a categorical variable, while dealing with
hist(full$Fare,xlab='Fare')#take log to make the data look more normal
```




```
hist(log(full$Fare+5),xlab="log(Fare)")
```




```
table(full$Cabin)
```

```
##
##                                   A10              A14              A16
```

```
##              1014              1              1              1 
##               A19            A20            A23            A24 
##                 1              1              1              1 
##               A26            A31            A32            A34 
##                 1              1              1              3 
##               A36             A5             A6             A7 
##                 1              1              1              1 
##              B101           B102            B18            B19 
##                 1              1              2              1 
##               B20            B22            B28             B3 
##                 2              2              2              1 
##               B30            B35            B37            B38 
##                 1              2              1              1 
##               B39             B4            B41            B42 
##                 1              1              2              1 
##               B49             B5            B50    B51 B53 B55 
##                 2              2              1              3 
## B57 B59 B63 B66        B58 B60            B69            B71 
##                 5              3              2              2 
##               B73            B77            B78            B79 
##                 1              2              2              1 
##               B80        B82 B84            B86            B94 
##                 1              1              1              1 
##           B96 B98           C101           C103           C104 
##                 4              3              1              1 
##              C106           C110           C111           C118 
##                 2              1              1              1 
##              C123           C124           C125           C126 
##                 2              2              2              2 
##              C128           C148             C2        C22 C26 
##                 1              1              2              4 
##       C23 C25 C27            C30            C32            C45 
##                 6              1              2              1 
##               C46            C47            C49            C50 
##                 2              1              1              1 
##               C52            C54        C62 C64            C65 
##                 2              2              2              2 
##               C68             C7            C70            C78 
##                 2              2              1              4 
##               C82            C83            C85            C86 
##                 1              2              2              2 
##               C87            C90            C91            C92 
##                 1              1              1              2 
##               C93            C95            C99              D 
##                 2              1              1              4 
##           D10 D12            D11            D15            D17 
##                 2              1              2              2 
##               D19            D20            D21            D26 
##                 2              2              2              2 
##               D28            D30            D33            D35 
##                 2              2              2              2 
##               D36            D37            D45            D46 
##                 2              2              1              1 
##               D47            D48            D49            D50 
```

```
##                 1                 1                 1                 1 
##               D56                D6                D7                D9 
##                 1                 1                 1                 1 
##               E10              E101               E12              E121 
##                 1                 3                 1                 2 
##               E17               E24               E25               E31 
##                 1                 2                 2                 2 
##               E33               E34               E36               E38 
##                 2                 3                 1                 1 
##               E40               E44               E46               E49 
##                 1                 2                 2                 1 
##               E50               E58               E63               E67 
##                 2                 1                 1                 2 
##               E68               E77                E8             F E69 
##                 1                 1                 2                 1 
##             F G63             F G73                F2               F33 
##                 2                 2                 4                 4 
##               F38                F4                G6                 T 
##                 1                 4                 5                 1 
##               A11               A18               A21               A29 
##                 1                 1                 1                 1 
##                A9               B10               B11               B24 
##                 1                 1                 1                 1 
##               B26               B36               B45       B52 B54 B56 
##                 1                 1                 2                 1 
##               B61              C105              C116              C130 
##                 1                 1                 2                 1 
##              C132               C28               C31               C39 
##                 1                 1                 2                 1 
##               C51               C53           C55 C57                C6 
##                 1                 1                 2                 2 
##               C80               C89               C97               D22 
##                 2                 2                 1                 1 
##               D34               D38               D40               D43 
##                 1                 1                 1                 1 
##           E39 E41               E45               E52               E60 
##                 1                 1                 1                 1 
##                 F             F E46             F E57 
##                 1                 1                 1
```

```
#Embarked seems to be irrelevant to survival status.


#data cleaning
#Convert PClass in to categorical data
full$PclassCat=as.factor(full$Pclass)
#Create categorical variable for SibSp and Parch
full$SibSpCat=factor(full$SibSp)
levels(full$SibSpCat)=list('0'=0,'1'=1,'>2'=c(2,3,4,5,8))
full$ParchCat=factor(full$Parch)
levels(full$ParchCat)=list('0'=0,'1'=1,'>2'=c(2,3,4,5,6,9))
#log Fare, median imputation for missing data
full$Fare[which(is.na(full$Fare))]=median(full$Fare,na.rm=T)
full$logFare=log(full$Fare+5)
```

```r
#Age, mean imputation for missing data
mu=mean(full$Age,na.rm=T)
full$Age2=full$Age
full$Age2[which(is.na(full$Age2))]=mu
#Split train data and test data
train=data.frame(Survived=train$Survived,full[1:ntrain,])#supervised label
test=data.frame(full[-(1:ntrain),])


#Models
colnames(train)
```

```
##  [1] "Survived"    "PassengerId" "Pclass"      "Name"        "Sex"
##  [6] "Age"         "SibSp"       "Parch"       "Ticket"      "Fare"
## [11] "Cabin"       "Embarked"    "PclassCat"   "SibSpCat"    "ParchCat"
## [16] "logFare"     "Age2"
```

```r
formulas=list(
  'Survived ~ PclassCat+Sex+Age2+SibSp+Parch+Fare',
'Survived ~ PclassCat+Sex+Age2+SibSpCat+ParchCat+Fare',
'Survived ~ PclassCat+Sex+Age2+SibSp+Parch+logFare',
'Survived ~ PclassCat+Sex+Age2+SibSpCat+ParchCat+logFare',
'Survived ~ Pclass+Sex+Age2+SibSp+Parch+Fare',
'Survived ~ Pclass+Sex+Age2+SibSpCat+ParchCat+Fare',
'Survived ~ Pclass+Sex+Age2+SibSp+Parch+logFare',
'Survived ~ Pclass+Sex+Age2+SibSpCat+ParchCat+logFare'
)


#5-fold Cross-Validation
bestThres=function(phat,y,thres.vec){
  minerr=Inf
  best=-1
  for(i in 1:length(thres.vec)){
    thres=thres.vec[i]
    pred=as.numeric(phat>thres)
    err=sum(pred!=y)
    if(err<minerr){
      minerr=err
      best=thres
    }
  }
  return(best)
}
```

```r
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4

## Loading required package: lattice

## Loading required package: ggplot2

## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018f.
## 1.0/zoneinfo/America/Los_Angeles'
```

```r
k=20
set.seed(0)
folds=createFolds(train$Survived,k=k)
thres.vec=seq(0,1,by=0.05)
errs=rep(0,length(formulas))
for(i in 1:length(folds)){
  for(j in 1:length(formulas)){
    mod=glm(formula=formulas[[j]],family=binomial,data=train[-folds[[i]],])
    phat=fitted(mod)
    thres=bestThres(phat,train$Survived[-folds[[i]]],thres.vec)
    phatpred=predict(mod,newdata=train[folds[[i]],],type='response')
    pred=as.numeric(phatpred>thres)
    err=sum(pred!=train$Survived[folds[[i]]])/length(folds[[i]])
    errs[j]=errs[j]+err
    }
}


errs=errs/length(folds)
errs
```

```
## [1] 0.1976515 0.1908081 0.1943182 0.1909091 0.1965404 0.1896717 0.1965909
## [8] 0.1998485
```

```r
bestid=which.min(errs)
bestid
```

```
## [1] 6
```

```r
formulas[[bestid]]
```

```
## [1] "Survived ~ Pclass+Sex+Age2+SibSpCat+ParchCat+Fare"
```

```r
#The best model chosed by cross validation is "Survived ~ Pclass+Sex+Age2+SibSpCat+ParchCat+Fare"
```

```r
#Prediction
mod=glm(formula=formulas[[bestid]],family=binomial,data=train)
summary(mod)
```

```
##
## Call:
## glm(formula = formulas[[bestid]], family = binomial, data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7248  -0.6312  -0.4321   0.6024   2.6838
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.804906   0.540108   8.896  < 2e-16 ***
## Pclass      -1.077924   0.138955  -7.757 8.67e-15 ***
## Sexmale     -2.714044   0.199378 -13.613  < 2e-16 ***
## Age2        -0.040093   0.007966  -5.033 4.82e-07 ***
## SibSpCat1    0.068130   0.221039   0.308  0.75791
## SibSpCat>2  -1.286655   0.384921  -3.343  0.00083 ***
## ParchCat1    0.336575   0.285628   1.178  0.23865
## ParchCat>2  -0.378336   0.318059  -1.190  0.23424
```

```
## Fare         0.002315   0.002283   1.014  0.31048
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  785.13  on 882  degrees of freedom
## AIC: 803.13
##
## Number of Fisher Scoring iterations: 5
```

```r
phat=fitted(mod)
thres=bestThres(phat,train$Survived,thres.vec)
phatpred=predict(mod,test,type='response')
    pred=as.numeric(phatpred>thres)
    sum(is.na(pred))
```

```
## [1] 0
```

```r
    pred.dat=data.frame(PassengerId=test$PassengerId,Survived=pred)
gender=read.csv('gender_submission.csv')

#Prediction accuracy
sum(gender$Survived==pred.dat$Survived)/418
```

```
## [1] 0.9354067
```

hw6.
stat151A.

2. In the logistic model, prove $Y-\hat{p}$ is orthogonal to column of X matrix.

· gradient of log-likelihood is $\nabla \ell(\beta) = X^T(Y-p)$.

$$\log \frac{p_i}{1-p_i} = x_i^T \beta.$$

thus the fitted probabilities satisfy $X^T(Y-\hat{p}) = 0$.

(a) Filling Missing values.

3. $z\text{-value} = \frac{\text{Estimate}}{\text{Std. Error}} = \frac{\tilde{\beta}}{se(\hat{\beta})}$

thus $s.e(\text{intercept}) = \frac{0.6864}{0.313} = 2.192971$

$se(\log(\text{distance})) = \frac{-0.9050}{-4.349} = 0.2080938$

} also the first 2 diagonal entries of $(X^T W X)^{-1}$.

null deviance:

null deviance is achieved under intercept model. In intercept model, $p_i = p$

$\sum_{i=1}^{n}(y_i \log p + (1-y_i)\log(1-p))$ the p which maximize the log-likelihood function is =

$$p = \bar{y}$$

know that $\bar{y} = 79/212$

substitute back into $\sum_{i=1}^{n}[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})] = 279.987$

$d.f = 212 - 1 = 211$ for null deviance.

Residual deviance = AIC − 2(# of explanatory variables + 1)

$= 222.18 - 2(3+1) = 214.18$.

$d.f = 212 - 3 - 1 = 208$ for residual deviance.

Missing value off diagonal of $(X^T W X)^{-1}$ can be obtained by symmetry, which is -0.2592180

Missing value on the 4th diagonal of $(X^T W X)^{-1}$ is $(0.3131)^2 = 0.09803162$.

(b). $X_0 = (\log(265), \log(26), 3.5)$ $\hat{\beta}$ can be obtained from (a)

thus $\hat{p}_0 = \frac{e^{x_0^T\hat{\beta}}}{1+e^{x_0^T\hat{\beta}}} = 0.7701945$

(c) By adding another variable, the residual deviance will decrease, just like RSS will decrease as variable increases in linear model. For larger model, we will have a likelihood less than or equal to maximized likelihood. Thus the residual deviance for larger models will be smaller.

The null deviance will not be affected because it is only depending on the intercept model.

$\boxed{4}$ $\hat{p} = \frac{e^{x_i^T \beta}}{1+e^{x^T \beta}}$

(a). log-likelihood can be written as:

$$\ell(\beta) = \sum_{i=1}^{n} [y_i(x_i^T\beta) - \log(1+\exp(x_i^T\beta))]$$

$$= Y^T X \cdot \beta - \sum_{i=1}^{n} \log(1+\exp(x_i^T\beta))$$

From the equation, we see that Y enters the log-likelihood only through $X^T Y$

(b). $\hat{p}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots \hat{\beta}_p x_{ip})}{1+\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots \hat{\beta}_p x_{ip})} = \frac{\exp(x_i^T\hat{\beta})}{1+\exp(x_i^T\hat{\beta})} = \frac{1}{1+\exp(-x_i^T\hat{\beta})}$

(c). $\sum_{i=1}^{n} y_i = y^T \cdot 1 = \hat{p}^T \cdot 1 = \sum_{i=1}^{n} \hat{p}_i$

y_is are 0 or 1. LHS will be the number of y_is which are equal to 1

(d). $\ell(\beta) = \sum_{i=1}^{n} [y_i(\log p_i) + (1-y_i)\log(1-p_i)]$

Residual Deviance: $-2\ell(\beta) = -2\sum_{i=1}^{n}[y_i(\log \hat{p}_i) + (1-y_i)(\log(1-\hat{p}_i)]$

where $\hat{p}_i = \frac{e^{x_i^T\hat{\beta}}}{1+e^{x_i^T\hat{\beta}}}$

$\boxed{5}$. (a). Missing z-value:

$4.11947/0.36342 = 11.33529$

Missing Coefficient:

$12.345 \cdot 0.028 = 0.34566$

$\bar{y} = 1813/4061$ $\quad n = 4601$

null deviance $= -2 \cdot n[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})] = 6170.153$

d.f. (of null deviance) $= n-1 = 4600$

d.f. residual deviance $= n-6-1 = 4594$

$AIC =$ residual deviance $+ 2(n+1) = 3245.1 + 2\cdot(6+1) = 3259.1$

(b). $x_0 = (\log(157+5), \log(0.868+5), \log(2.894+5), \log(5), \log(5), \log(5))$.

$\hat{p}_0 = \frac{\exp(x_0^T\beta)}{1+\exp(x_0^T\beta)} = 0.958115$

(c). M1 has higher maximum likelihood than M2, M1 is preferrable.
Also, M1 has lower AIC than M2. In this sense M1 is also preferrable.