

STAT 381 Final Project

Zachary Lazerick, Sebastian Johns, Kelly Zhou

2022-04-30

```
library(readr)
Income <- read.csv("income_evaluation.csv", na.strings = " ?")
Income <- na.omit(Income)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union
```

```
attach(Income)
Income$sex <- as.factor(Income$sex)
Income$income <- as.factor(Income$income)
Income$workclass <- as.factor(Income$workclass)
Income$education <- as.factor(Income$education)
Income$marital.status <- as.factor(Income$marital.status)
Income$occupation <- as.factor(Income$occupation)
Income$race <- as.factor(Income$race)
Income$relationship <- as.factor(Income$relationship)
Income$native.country <- as.factor(Income$native.country)
Income$native.country <- recode(Income$native.country, " Cambodia" = "E-AS", " Canada" = "NA", " China" = "NA")
summary(Income)
```

```
##           age           workclass           fnlwtg
## Min.      :17.00   Federal-gov      : 943   Min.      : 13769
## 1st Qu.:28.00   Local-gov        : 2067   1st Qu.: 117627
## Median :37.00   Private          :22286   Median : 178425
## Mean    :38.44   Self-emp-inc     : 1074   Mean    : 189794
## 3rd Qu.:47.00   Self-emp-not-inc: 2499   3rd Qu.: 237628
## Max.    :90.00   State-gov        : 1279   Max.    :1484705
##           Without-pay      : 14
##           education education.num           marital.status
## HS-grad      :9840   Min.      : 1.00   Divorced      : 4214
## Some-college:6678   1st Qu.: 9.00   Married-AF-spouse : 21
## Bachelors    :5044   Median :10.00   Married-civ-spouse :14065
## Masters      :1627   Mean    :10.12   Married-spouse-absent: 370
## Assoc-voc    :1307   3rd Qu.:13.00   Never-married   : 9726
## 11th         :1048   Max.    :16.00   Separated       : 939
```

```

## (Other) :4618 Widowed : 827
## occupation relationship race
## Prof-specialty :4038 Husband :12463 Amer-Indian-Eskimo: 286
## Craft-repair :4030 Not-in-family : 7726 Asian-Pac-Islander: 895
## Exec-managerial:3992 Other-relative: 889 Black : 2817
## Adm-clerical :3721 Own-child : 4466 Other : 231
## Sales :3584 Unmarried : 3212 White :25933
## Other-service :3212 Wife : 1406
## (Other) :7585
## sex capital.gain capital.loss hours.per.week
## Female: 9782 Min. : 0 Min. : 0.00 Min. : 1.00
## Male :20380 1st Qu.: 0 1st Qu.: 0.00 1st Qu.:40.00
## Median : 0 Median : 0.00 Median :40.00
## Mean : 1092 Mean : 88.37 Mean :40.93
## 3rd Qu.: 0 3rd Qu.: 0.00 3rd Qu.:45.00
## Max. :99999 Max. :4356.00 Max. :99.00
##
## native.country income
## E-AS: 663 <=50K:22654
## NA : 717 >50K : 7508
## SA : 86
## CA : 534
## EU : 493
## ME : 42
## US :27627

```

```
cor(age, education.num)
```

```
## [1] 0.04352609
```

```
cor(age,fnlwgt)
```

```
## [1] -0.07651084
```

```
cor(age, hours.per.week)
```

```
## [1] 0.1015988
```

```
cor(age,capital.gain)
```

```
## [1] 0.08015423
```

```
cor(age,capital.loss)
```

```
## [1] 0.06016548
```

```
cor(education.num,fnlwgt)
```

```
## [1] -0.04499174
```

```
cor(education.num,capital.gain)
```

```
## [1] 0.124416
```

```
cor(education.num,capital.loss)
```

```
## [1] 0.07964641
```

```
cor(education.num,hours.per.week)
```

```
## [1] 0.1525221
cor(fnlwgt, capital.gain)

## [1] 0.0004215674
cor(fnlwgt, capital.loss)

## [1] -0.009749528
cor(fnlwgt, hours.per.week)

## [1] -0.02288575
cor(hours.per.week, capital.gain)

## [1] 0.0804318
cor(hours.per.week, capital.loss)

## [1] 0.05241705
cor(capital.gain, capital.loss)

## [1] -0.03222933
```

there appears to be no correlation among the numerical variables. However, we can assume correlation between marital status and relationship status. There is also obvious correlation between education and education number as well as workclass and occupation.

```
library(leaps)
regfit.full <- regsubsets(income ~ native.country + hours.per.week + sex + race + relationship + marital.status, data = d)

reg.summary <- summary(regfit.full)
names(reg.summary)

## [1] "which" "rssq" "rss" "adjr2" "cp" "bic" "outmat" "obj"
reg.summary$adjr2

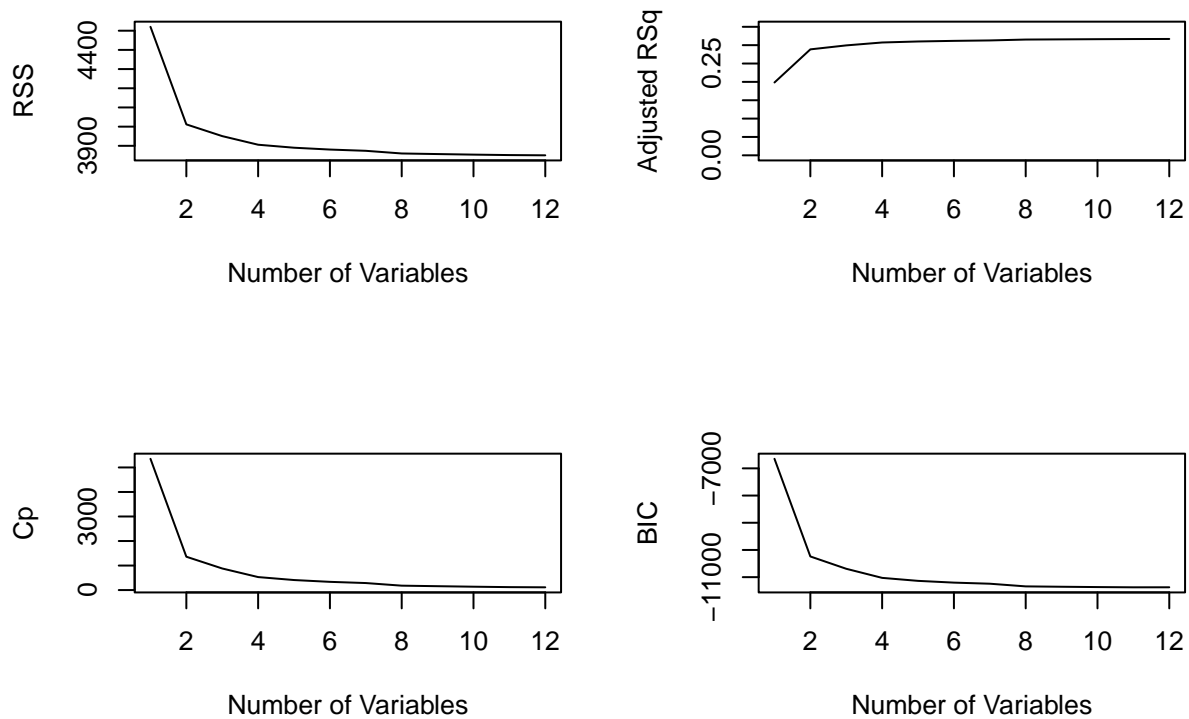
## [1] 0.1983703 0.2885327 0.2993559 0.3072913 0.3100308 0.3117202 0.3128546
## [8] 0.3152942 0.3158062 0.3162939 0.3166976 0.3168935

par(mfrow = c(2,2))
plot(reg.summary$rss, xlab = "Number of Variables", ylab = "RSS", type = "l")

plot(reg.summary$adjr2, xlab = "Number of Variables", ylab = "Adjusted RSq", type = "l", ylim=c(0,.35))

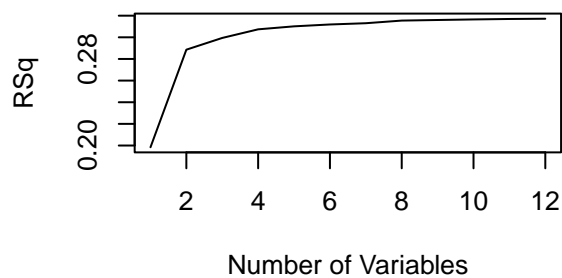
plot(reg.summary$cp, xlab = "Number of Variables", ylab = "Cp", type = "l")

plot(reg.summary$bic, xlab = "Number of Variables", ylab = "BIC", type = "l")
```



```
plot(reg.summary$rsq, xlab = "Number of Variables", ylab = "RSq", type = "l")
coef(regfit.full, 4)
```

```
##                (Intercept)                hours.per.week
##                0.336432905                0.003337259
## marital.status Married-civ-spouse        education.num
##                0.319635395                0.048628172
##                age
##                0.003503305
```



```
reg.summary$adjr2
```

```
## [1] 0.1983703 0.2885327 0.2993559 0.3072913 0.3100308 0.3117202 0.3128546
## [8] 0.3152942 0.3158062 0.3162939 0.3166976 0.3168935
```

```
reg.summary$cp
```

```
## [1] 5352.0157 1359.0202 880.5679 530.0572 409.7070 335.8662 286.6179
## [8] 179.5655 157.8830 137.2810 120.4013 112.7250
```

```
reg.summary$bic
```

```
## [1] -6649.445 -10238.981 -10692.033 -11026.279 -11136.482 -11201.114
## [7] -11241.550 -11339.512 -11352.763 -11364.957 -11373.456 -11372.789
```

```

set.seed(1)
train <- sample(30162,30162*.7)
Income.test <- Income[-train,]
dim(Income.test)

## [1] 9049 15

income.test <- income[-train]
glm.fit <- glm(income~age+education.num+hours.per.week+marital.status, data = Income, family = binomial)
glm.probs <- predict(glm.fit,Income.test,type = "response")
glm.pred <- rep("<=50k",9049)
glm.pred[glm.probs > .5] <- ">50k"
table(glm.pred,income.test)

##          income.test
## glm.pred  <=50K  >50K
##    =<50k    6311  1088
##    >50k     554  1096

(6311+1096)/9049

## [1] 0.8185435

summary(glm.fit)

##
## Call:
## glm(formula = income ~ age + education.num + hours.per.week +
##      marital.status, family = binomial, data = Income, subset = train)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.8237  -0.6079  -0.2691   0.3713   3.3711
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -9.008241    0.170595 -52.805 < 2e-16 ***
## age             0.031430    0.001776  17.693 < 2e-16 ***
## education.num   0.385912    0.009005  42.856 < 2e-16 ***
## hours.per.week  0.031627    0.001773  17.842 < 2e-16 ***
## marital.status Married-AF-spouse  2.490356    0.542567   4.590 4.43e-06 ***
## marital.status Married-civ-spouse  2.067468    0.066910  30.899 < 2e-16 ***
## marital.status Married-spouse-absent -0.284559    0.257524  -1.105 0.269
## marital.status Never-married    -0.516309    0.088638  -5.825 5.72e-09 ***
## marital.status Separated        -0.075085    0.165752  -0.453 0.651
## marital.status Widowed          -0.071196    0.163096  -0.437 0.662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 23845  on 21112  degrees of freedom
## Residual deviance: 16091  on 21103  degrees of freedom
## AIC: 16111
##
## Number of Fisher Scoring iterations: 6

```

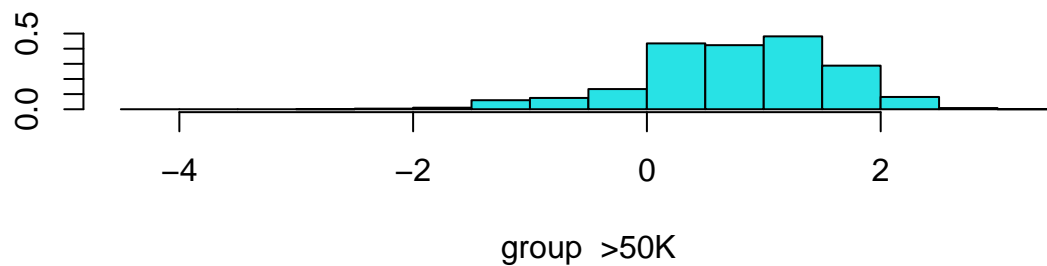
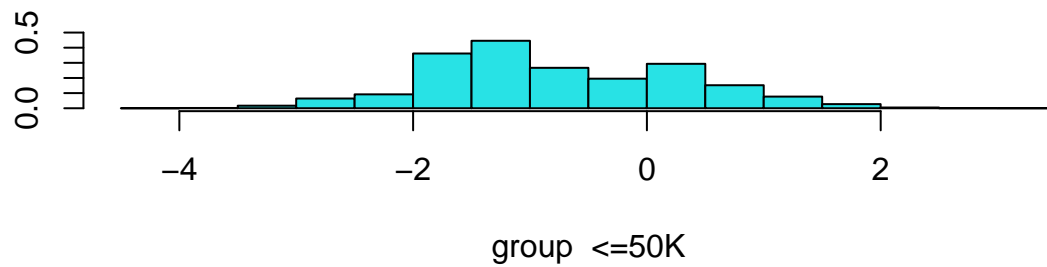
LDA

```
library(MASS)

##
## Attaching package: 'MASS'
## The following object is masked from 'package:dplyr':
##
##      select
lda.fit <- lda(income~age+hours.per.week+education.num+marital.status, data = Income, subset = train)
lda.fit

## Call:
## lda(income ~ age + hours.per.week + education.num + marital.status,
##      data = Income, subset = train)
##
## Prior probabilities of groups:
##      <=50K      >50K
## 0.7478331 0.2521669
##
## Group means:
##      age hours.per.week education.num marital.status Married-AF-spouse
## <=50K 36.67040      39.41041      9.631452      0.0005700171
## >50K 44.01803      45.72502     11.610443      0.0013148009
##      marital.status Married-civ-spouse marital.status Married-spouse-absent
## <=50K      0.3396035      0.015517132
## >50K      0.8518032      0.003756574
##      marital.status Never-married marital.status Separated
## <=50K      0.40623219      0.037747799
## >50K      0.06104433      0.009579264
##      marital.status Widowed
## <=50K      0.03166762
## >50K      0.01070624
##
## Coefficients of linear discriminants:
##                                LD1
## age                        0.01819334
## hours.per.week             0.01665735
## education.num              0.24400337
## marital.status Married-AF-spouse 1.81925399
## marital.status Married-civ-spouse 1.62835507
## marital.status Married-spouse-absent 0.12185349
## marital.status Never-married 0.03913206
## marital.status Separated 0.07304818
## marital.status Widowed 0.01467706

plot(lda.fit)
```



```
lda.pred <- predict(lda.fit,Income.test)
names(lda.pred)
```

```
## [1] "class"      "posterior" "x"
```

```
lda.class <- lda.pred$class
table(lda.class,income.test)
```

```
##           income.test
## lda.class  <=50K >50K
##    <=50K    6264 1050
##    >50K      601 1134
```

```
mean(lda.class == income.test)
```

```
## [1] 0.8175489
```