471 mid

jiajian huang

March 7, 2019

```
library(knitr)
library(rmdformats)
library(arm)
library(leaps)
library(tableone)
library(pander)
library(MASS)
library(ROCR)
library(skimr)
library(rms)
library(broom)
library(dplyr)
library(tidyverse)
```

Load and Tidy the Data

```
test <- read.csv("census_test.csv")
train <- read.csv("census_train.csv")
```

```
skim(train)
```

	variable <chr>	type <chr>	stat <chr>	level <chr>
1	age	integer	missing	.all
2	age	integer	complete	.all
3	age	integer	n	.all
4	age	integer	mean	.all
5	age	integer	sd	.all
6	age	integer	p0	.all
7	age	integer	p25	.all
8	age	integer	p50	.all
9	age	integer	p75	.all
10	age	integer	p100	.all

1-10 of 224 rows | 1-5 of 7 columns Previous **1** 2 3 4 5 6 ... 23 Next

```
skim(test)
```

	variable <chr>	type <chr>	stat <chr>	level <chr>
1	age	integer	missing	.all
2	age	integer	complete	.all
3	age	integer	n	.all
4	age	integer	mean	.all
5	age	integer	sd	.all
6	age	integer	p0	.all
7	age	integer	p25	.all
8	age	integer	p50	.all
9	age	integer	p75	.all
10	age	integer	p100	.all

1-10 of 222 rows | 1-5 of 7 columns   Previous **1** 2 3 4 5 6 ... 23 Next

```
names(train)
```

```
 [1] "age"            "workclass"      "fnlwgt"         "education"
 [5] "education.num"  "marital.status" "occupation"     "relationship"
 [9] "race"           "sex"            "capital.gain"   "capital.loss"
[13] "hours.per.week" "native.country" "income"
```

```
dim(train)
```

```
[1] 25000    15
```

There are 250000 subjects in the [redacted] data set and 15 variables for each subjects.There are none missing values for any one of the variables.

Check categorical variables

```
levels(train$native.country)
```

```
 [1] " ?"                          " Cambodia"
 [3] " Canada"                     " China"
 [5] " Columbia"                   " Cuba"
 [7] " Dominican-Republic"         " Ecuador"
 [9] " El-Salvador"                " England"
```

```
[11] " France"                      " Germany"
[13] " Greece"                      " Guatemala"
[15] " Haiti"                       " Holand-Netherlands"
[17] " Honduras"                    " Hong"
[19] " Hungary"                     " India"
[21] " Iran"                        " Ireland"
[23] " Italy"                       " Jamaica"
[25] " Japan"                       " Laos"
[27] " Mexico"                      " Nicaragua"
[29] " Outlying-US(Guam-USVI-etc)" " Peru"
[31] " Philippines"                 " Poland"
[33] " Portugal"                    " Puerto-Rico"
[35] " Scotland"                    " South"
[37] " Taiwan"                      " Thailand"
[39] " Trinadad&Tobago"             " United-States"
[41] " Vietnam"                     " Yugoslavia"
```

```
library(rockchalk)
train$native.country <- combineLevels(train$native.country,levs = c(" ?"," Cambodia"," Canada"
," China"," Columbia"," Cuba",                       " Dominican-Republic"," Ecuador"," El-Sal
vador"," England"," France"," Germany" ," Greece"," Guatemala" , " Haiti",
  " Holand-Netherlands" ," Honduras"," Hong"," Hungary"," India"," Iran"," Ireland"," Italy" ,
 " Jamaica" ," Japan",                        " Laos"," Mexico"," Nicaragua"," Outlying-US(Guam-
USVI-etc)"," Peru"," Philippines"," Poland" ," Portugal" ,
                                                                  " Puerto-Rico"," Scotland"
," South"," Taiwan" , " Thailand",
                                                                  " Trinadad&Tobago"," Vietn
am", " Yugoslavia"),
                                      newLabel = c("Non-US") )
```

```
The original levels  ?  Cambodia  Canada  China  Columbia  Cuba  Dominican-Republic  Ecuador
El-Salvador  England  France  Germany  Greece  Guatemala  Haiti  Holand-Netherlands  Honduras
 Hong  Hungary  India  Iran  Ireland  Italy  Jamaica  Japan  Laos  Mexico  Nicaragua  Outlying
-US(Guam-USVI-etc)  Peru  Philippines  Poland  Portugal  Puerto-Rico  Scotland  South  Taiwan
 Thailand  Trinadad&Tobago  United-States  Vietnam  Yugoslavia
have been replaced by  United-States Non-US
```

```
levels(train$education)
```

```
 [1] " 10th"          " 11th"          " 12th"          " 1st-4th"
 [5] " 5th-6th"       " 7th-8th"       " 9th"           " Assoc-acdm"
 [9] " Assoc-voc"     " Bachelors"     " Doctorate"     " HS-grad"
[13] " Masters"       " Preschool"     " Prof-school"   " Some-college"
```

```
train$education <- combineLevels(train$education,levs = c(" 10th"," 11th"," 12th"," 1st-4th","
 5th-6th"," 7th-8th"," 9th"," Preschool",
                                                              " HS-grad"),newLabel = c("High Schoo
l and below") )
```

```
The original levels  10th  11th  12th  1st-4th  5th-6th  7th-8th  9th  Assoc-acdm  Assoc-voc
```

```
Bachelors  Doctorate  HS-grad  Masters  Preschool  Prof-school  Some-college
have been replaced by  Assoc-acdm  Assoc-voc  Bachelors  Doctorate  Masters  Prof-school  Some
-college High School and below
```

```
train$education <- combineLevels(train$education,levs = c(" Assoc-acdm"," Assoc-voc"," Some-co
llege" ),newLabel = c("some college") )
```

```
The original levels  Assoc-acdm  Assoc-voc  Bachelors  Doctorate  Masters  Prof-school  Some-c
ollege High School and below
have been replaced by  Bachelors  Doctorate  Masters  Prof-school High School and below some c
ollege
```

```
train$education <- combineLevels(train$education,levs = c(" Bachelors"," Doctorate"," Masters"
," Prof-school"),newLabel = c("Bachelors and above") )
```

```
The original levels  Bachelors  Doctorate  Masters  Prof-school High School and below some col
lege
have been replaced by High School and below some college Bachelors and above
```

```
train$income <- combineLevels(train$income,levs = c(" <=50K"),newLabel = c("0") )
```

```
The original levels  <=50K  >50K
have been replaced by  >50K 0
```

```
train$income <- combineLevels(train$income,levs = c(" >50K"),newLabel = c("1") )
```

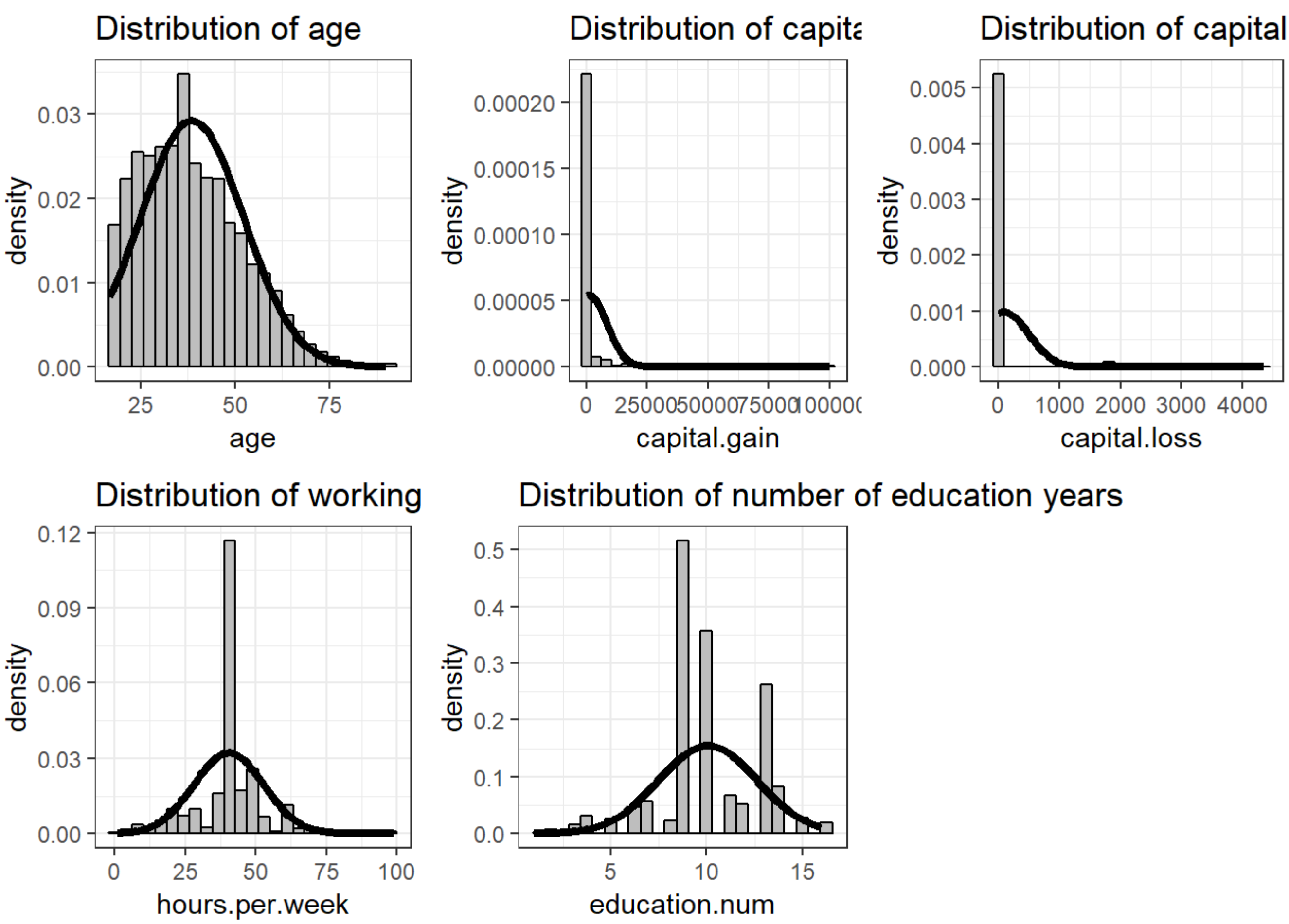
```
The original levels  >50K 0
have been replaced by 0 1
```

First, we want to check if it is necessary for some of the categorical variables to combine some levels together. We combine the variable `native.country` to two category `United States` and `Non-US`. Also, we combine 16 levels in the variable `education` into 3 categories `High School and below`, `some college`, and `Bachelors and above`. In order to build our logistic model, we re code levels of `income` into **0** and **1** where **0** stands for income less than or equal to 50K and **1** stands for income more than 50K.

Quantative variables

```
p1<- ggplot(train, aes(x = age)) + geom_histogram(aes(y = ..density..),bins=25, color = "black
", fill = "grey") +
  stat_function(fun = dnorm,args = list(mean = mean(train$age), sd = sd(train$age)),
               lwd = 1.5, col = "black") +
  theme_bw()+labs(title = "Distribution of age")
p2<- ggplot(train, aes(x =  capital.gain)) + geom_histogram(aes(y = ..density..),bins=25, colo
r = "black", fill = "grey") +
  stat_function(fun = dnorm,args = list(mean = mean(train$capital.gain), sd = sd(train$ capita
l.gain)),
               lwd = 1.5, col = "black") +
```

```
  theme_bw()+labs(title = "Distribution of capital gain")
p3<- ggplot(train, aes(x =  capital.loss)) + geom_histogram(aes(y = ..density..),bins=25, colo
r = "black", fill = "grey") +
  stat_function(fun = dnorm,args = list(mean = mean(train$capital.loss), sd = sd(train$capital
.loss)),
                lwd = 1.5, col = "black") +
  theme_bw()+labs(title = "Distribution of capital loss")
p4<- ggplot(train, aes(x =  hours.per.week)) + geom_histogram(aes(y = ..density..),bins=25, co
lor = "black", fill = "grey") +
  stat_function(fun = dnorm,args = list(mean = mean(train$ hours.per.week), sd = sd(train$ hou
rs.per.week)),
                lwd = 1.5, col = "black") +
  theme_bw()+labs(title = "Distribution of working hours per week ")
p5<- ggplot(train, aes(x = education.num)) + geom_histogram(aes(y = ..density..),bins=25, colo
r = "black", fill = "grey") +
  stat_function(fun = dnorm,args = list(mean = mean(train$education.num), sd = sd(train$educat
ion.num)),
                lwd = 1.5, col = "black") +
  theme_bw()+labs(title = "Distribution of number of education years ")
gridExtra::grid.arrange(p1, p2, p3, p4,p5, nrow = 2)
```



The histograms show that the quantitative variables are normally distributed except the captial gain and loss, which includes a lot of value 0. We may consider spend more degree of freedom for these two variables latter in the model fitting process.

Test dataset

We did the same treatment to test dataset as to train dataset.

```
test$education <- combineLevels(test$education,levs = c(" 10th"," 11th"," 12th"," 1st-4th"," 5
th-6th"," 7th-8th"," 9th"," Preschool",
                                                                   " HS-grad"),newLabel = c("High Schoo
l and below") )
```

```
The original levels  10th  11th  12th  1st-4th  5th-6th  7th-8th  9th  Assoc-acdm  Assoc-voc
Bachelors  Doctorate  HS-grad  Masters  Preschool  Prof-school  Some-college
have been replaced by  Assoc-acdm  Assoc-voc  Bachelors  Doctorate  Masters  Prof-school  Some
-college High School and below
```

```
test$education <- combineLevels(test$education,levs = c(" Assoc-acdm"," Assoc-voc"," Some-coll
ege" ),newLabel = c("some college") )
```

```
The original levels  Assoc-acdm  Assoc-voc  Bachelors  Doctorate  Masters  Prof-school  Some-c
ollege High School and below
have been replaced by  Bachelors  Doctorate  Masters  Prof-school High School and below some c
ollege
```

```
test$education <- combineLevels(test$education,levs = c(" Bachelors"," Doctorate"," Masters","
 Prof-school"),newLabel = c("Bachelors and above") )
```

```
The original levels  Bachelors  Doctorate  Masters  Prof-school High School and below some col
lege
have been replaced by High School and below some college Bachelors and above
```

```
test$income <- combineLevels(test$income,levs = c(" <=50K"),newLabel = c("0") )
```

```
The original levels  <=50K  >50K
have been replaced by  >50K 0
```

```
test$income <- combineLevels(test$income,levs = c(" >50K"),newLabel = c("1") )
```

```
The original levels  >50K 0
have been replaced by 0 1
```

```
test$native.country <- combineLevels(test$native.country,levs = c(" ?"," Cambodia"," Canada","
 China"," Columbia",
                                                                          " Cuba"," Dominican-Republ
ic"," Ecuador"," El-Salvador",
                                                                          " England" ," France"," Ge
rmany"," Greece", " Guatemala",
                                                                          " Haiti"," Honduras"," Hon
```

```
g"," Hungary"," India",
                                                                     " Iran"," Ireland"," Italy
"," Jamaica"," Japan",
                                                                     " Laos" ," Mexico"," Nicar
agua", " Outlying-US(Guam-USVI-etc)",
                                                                     " Peru"," Philippines"," P
oland"," Portugal",
                                                                     " Puerto-Rico"," Scotland"
," South"," Taiwan",
                                                                     " Thailand"," Trinadad&Tob
ago"," Vietnam",
                                                                     " Yugoslavia"),
                                       newLabel = c("Non-US") )
```

```
The original levels  ?  Cambodia  Canada  China  Columbia  Cuba  Dominican-Republic  Ecuador
El-Salvador  England  France  Germany  Greece  Guatemala  Haiti  Honduras  Hong  Hungary  Indi
a  Iran  Ireland  Italy  Jamaica  Japan  Laos  Mexico  Nicaragua  Outlying-US(Guam-USVI-etc)
Peru  Philippines  Poland  Portugal  Puerto-Rico  Scotland  South  Taiwan  Thailand  Trinadad&
Tobago  United-States  Vietnam  Yugoslavia
have been replaced by  United-States Non-US
```

Codebook

```
a <- dput(names(train))
```

```
c("age", "workclass", "fnlwgt", "education", "education.num",
"marital.status", "occupation", "relationship", "race", "sex",
"capital.gain", "capital.loss", "hours.per.week", "native.country",
"income")
```

```
options(width = 200)
b <- c("age at baseline",
       "Working class",
       "The final sample weight",
       "Education level",
       "Number of years spent on education",
       "Marriage status",
       "Ocupation",
       "Role in family",
       "Race",
       "Sex",
       "Capital gain in a year",
       "Capital loss in a year",
       "hours spent on work per week",
       "Country born in",
       "Total income")
c <- map(train, function(x) class(x))
d <- map(train, function(x) sum(is.na(x)))
e <- map(train, function(x) ifelse(is.factor(x) == T, "--", min(x, na.rm=T)))
f <- map(train, function(x) ifelse(is.factor(x) == T, "--", max(x, na.rm=T)))
```

```
train.CB <- data_frame(Variable = a, Description = b, Class = c, Missing = d, Min = e, Max = f
)
pander(train.CB)
```

Table continues below

Variable	Description	Class	Missing	Min
age	age at baseline	integer	0	17
workclass	Working class	factor	0	–
fnlwgt	The final sample weight	integer	0	12285
education	Education level	factor	0	–
education.num	Number of years spent on education	integer	0	1
marital.status	Marriage status	factor	0	–
occupation	Ocupation	factor	0	–
relationship	Role in family	factor	0	–
race	Race	factor	0	–
sex	Sex	factor	0	–
capital.gain	Capital gain in a year	integer	0	0
capital.loss	Capital loss in a year	integer	0	0
hours.per.week	hours spent on work per week	integer	0	1
native.country	Country born in	factor	0	–
income	Total income	factor	0	–

Max
90
1484705
–
16
–
–
–
–
–
99999
4356
99
–

```
rm(a, b, c, d, e)
```

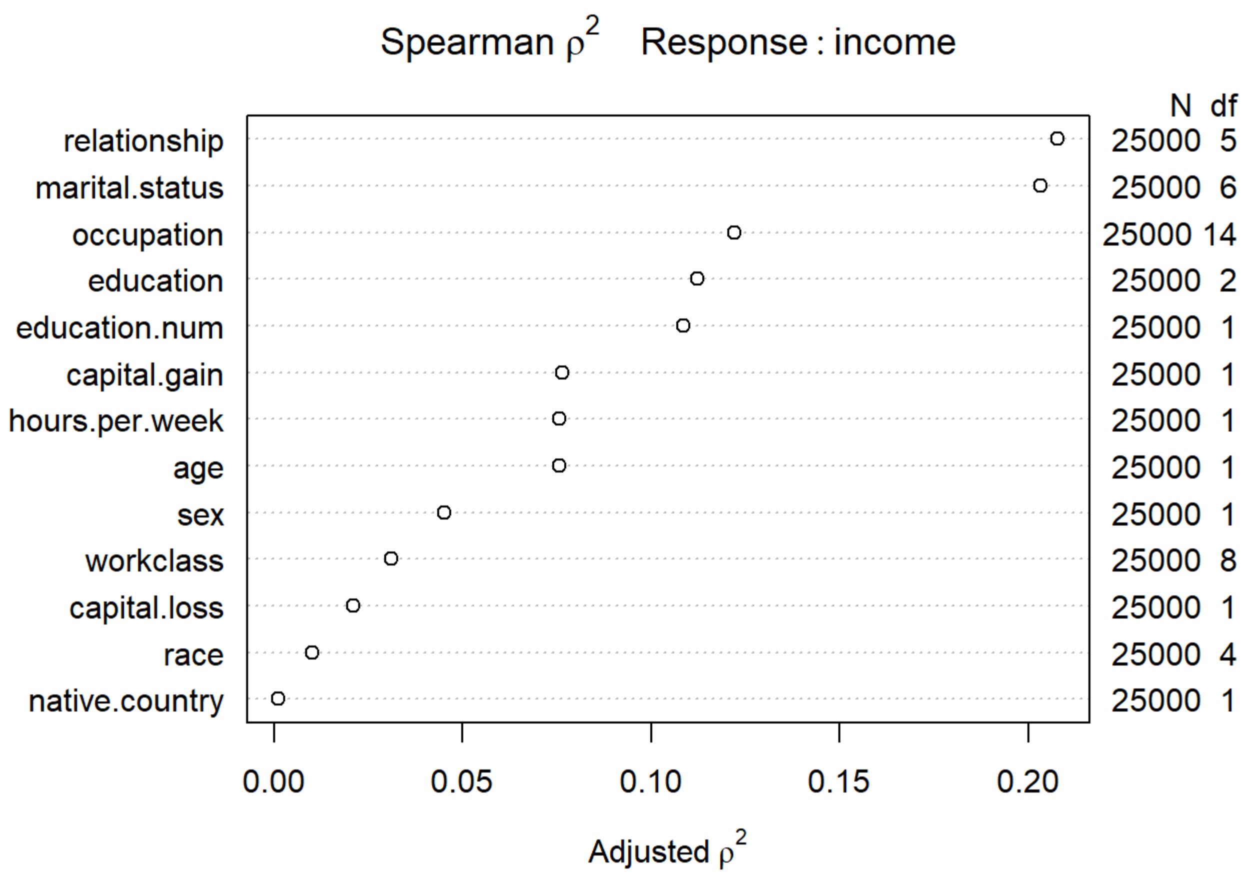
Losgitc model

```
names(train)
```

```
 [1] "age"              "workclass"       "fnlwgt"          "education"       "education.num"   "mar
ital.status" "occupation"     "relationship"    "race"            "sex"             "capital.gain
"
[12] "capital.loss"    "hours.per.week" "native.country" "income"
```

We'll start with a model motivated by the Spearman ρ2 plot developed above, and repeated below.

```
plot(spearman2(income ~ age+workclass+education+education.num+marital.status+occupation+relati
onship+race+sex
               +capital.gain+capital.loss+hours.per.week+native.country,
               data = train))
```



First, we try to use best subsets to select predcitors.

Running "Best Subsets" to select predictors

```
preds <- with(train, cbind(age,workclass,education,education.num,marital.status,occupation,rel
ationship,race,sex,
                           capital.gain,capital.loss,hours.per.week,native.country))
```

```r
x1 <- regsubsets(preds, train$income, nvmax=13)
rs.sum <- summary(x1)
rs.sum
```

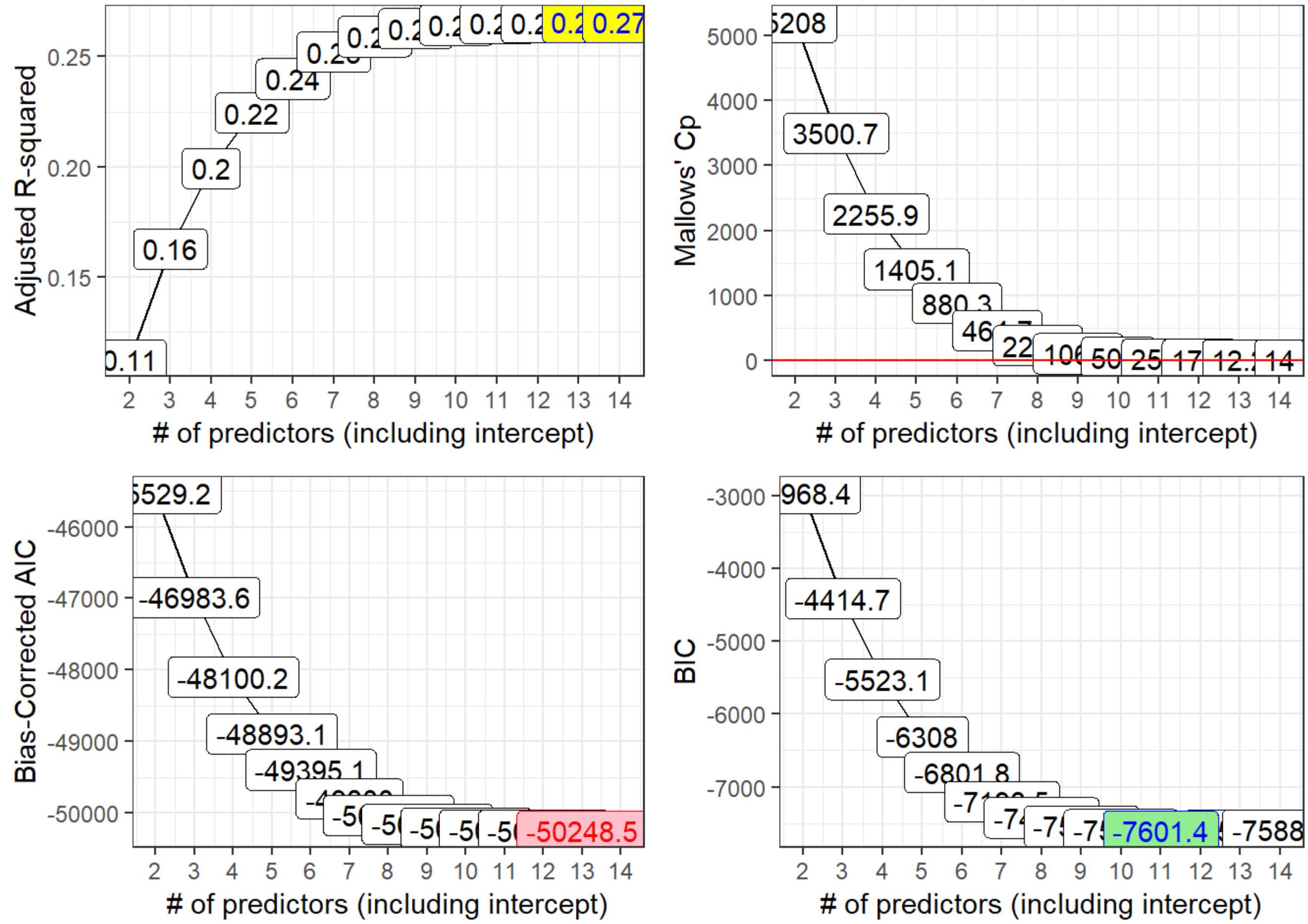
```
Subset selection object
13 Variables  (and intercept)
               Forced in Forced out
age                FALSE      FALSE
workclass          FALSE      FALSE
education          FALSE      FALSE
education.num      FALSE      FALSE
marital.status     FALSE      FALSE
occupation         FALSE      FALSE
relationship       FALSE      FALSE
race               FALSE      FALSE
sex                FALSE      FALSE
capital.gain       FALSE      FALSE
capital.loss       FALSE      FALSE
hours.per.week     FALSE      FALSE
native.country     FALSE      FALSE
1 subsets of each size up to 13
Selection Algorithm: exhaustive
          age workclass education education.num marital.status occupation relationship race se
x capital.gain capital.loss hours.per.week native.country
1  ( 1 )  " " " "       " "       "*"           " "            " "        " "          " "  "
" " "          " "          " "            " "           
2  ( 1 )  "*" " "       " "       "*"           " "            " "        " "          " "  "
" " "          " "          " "            " "           
3  ( 1 )  "*" " "       " "       "*"           " "            " "        " "          " "  "*
" " "          " "          " "            " "           
4  ( 1 )  "*" " "       " "       "*"           " "            " "        " "          " "  "*
" "*"          " "          " "            " "           
5  ( 1 )  "*" " "       " "       "*"           " "            " "        " "          " "  "*
" "*"          " "          "*"            " "           
6  ( 1 )  "*" " "       " "       "*"           " "            " "        " "          " "  "*
" "*"          "*"          "*"            " "           
7  ( 1 )  "*" " "       " "       "*"           "*"            " "        " "          " "  "*
" "*"          "*"          "*"            " "           
8  ( 1 )  "*" " "       "*"       "*"           "*"            " "        " "          " "  "*
" "*"          "*"          "*"            " "           
9  ( 1 )  "*" " "       "*"       "*"           "*"            " "        "*"          " "  "*
" "*"          "*"          "*"            " "           
10  ( 1 ) "*" " "       "*"       "*"           "*"            " "        "*"          "*"  "*
" "*"          "*"          "*"            " "           
11  ( 1 ) "*" " "       "*"       "*"           "*"            "*"        "*"          "*"  "*
" "*"          "*"          "*"            " "           
12  ( 1 ) "*" "*"       "*"       "*"           "*"            "*"        "*"          "*"  "*
" "*"          "*"          "*"            " "           
13  ( 1 ) "*" "*"       "*"       "*"           "*"            "*"        "*"          "*"  "*
" "*"          "*"          "*"            "*"           
```

```r
rs.sum$adjr2<-round(rs.sum$adjr2, 4)
rs.sum$cp<-round(rs.sum$cp, 1)
```

```
rs.sum$bic<-round(rs.sum$bic, 1)
rs.sum$aic.corr <- 25000*log(rs.sum$rss / 25000) + 2*(2:14) +
               (2 * (2:14) * ((2:14)+1) / (25000 - (2:14) - 1))
rs.sum$aic.corr<-round(rs.sum$aic.corr,1)
```

```
best_mods_1 <- data_frame( k = 2:14,
                           r2 = rs.sum$rsq, adjr2 = rs.sum$adjr2, cp = rs.sum$cp, aic.c = rs.s
um$aic.c, bic = rs.sum$bic)
rs.sum <- cbind(best_mods_1, rs.sum$which)
```

```
p1<- ggplot(rs.sum, aes(x = k, y = adjr2,label = round(adjr2,2))) +
  geom_line() +
  geom_label() +
  geom_label(data = subset(rs.sum,adjr2 == max(adjr2)),aes(x = k, y = adjr2, label = round(adj
r2,2)),
             fill = "yellow", col = "blue") + theme_bw() +
  scale_x_continuous(breaks = 2:14) +
  labs(x = "# of predictors (including intercept)",y = "Adjusted R-squared")
p2<- ggplot(rs.sum, aes(x = k, y = cp,
                   label = round(cp,1))) +
  geom_line() +geom_label() +geom_abline(intercept = 0, slope = 1,col = "red") + theme_bw() +
  scale_x_continuous(breaks = 2:14) +
  labs(x = "# of predictors (including intercept)",y = "Mallows' Cp")
p3<- ggplot(rs.sum, aes(x = k, y = aic.c,label = round(aic.c,1))) +geom_line() +
  geom_label() +geom_label(data = subset(rs.sum, aic.c == min(aic.c)),aes(x = k, y = aic.c), f
ill = "pink",
                           col = "red") + theme_bw() +
  scale_x_continuous(breaks = 2:14) +labs(x = "# of predictors (including intercept)",y = "Bia
s-Corrected AIC")
p4<- ggplot(rs.sum, aes(x = k, y = bic,label = round(bic,1))) +
  geom_line() +
  geom_label() +
  geom_label(data = subset(rs.sum, bic == min(bic)),aes(x = k, y = bic),fill = "lightgreen", c
ol = "blue") + theme_bw() +
  scale_x_continuous(breaks = 2:14) +
  labs(x = "# of predictors (including intercept)",y = "BIC")
gridExtra::grid.arrange(p1, p2, p3, p4, nrow = 2)
```



Candidate Models from Best Subsets

The models we'll consider are:

Inputs	Predictors Included	Reason
10	`age` `education` `education.num` `marital.status` `relationship` `race` `sex` `capital.gain` `capital.loss` `hours.per.week`	lowest BIC
9	`age` `education` `education.num` `marital.status` `relationship` `sex` `capital.gain` `capital.loss` `hours.per.week`	suggested by C_p
12	`age` `education` `education.num` `marital.status` `relationship` `race` `sex` `capital.gain` `capital.loss` `hours.per.week` `workclass` `occupation`	lowest AIC (corr.)
12	`age` `education` `education.num` `marital.status` `relationship` `race` `sex` `capital.gain` `capital.loss` `hours.per.week` `workclass` `occupation`	highest adj. R^2

```
glm9 <- glm(income~age+education+education.num+marital.status+relationship+sex+capital.gain+capital.loss+hours.per.week,data = train,family = binomial)
```

```
glm10 <- glm(income~age+education+education.num+marital.status+relationship+sex+capital.gain+c
apital.loss+hours.per.week+race,data = train,family = binomial)
glm12 <- glm(income~age+education+education.num+marital.status+relationship+sex+capital.gain+c
apital.loss+hours.per.week+occupation
             +race+workclass,data = train,family = binomial)
glm0<-glm(income~1, data = train, family = binomial)
```

```
anova(glm12,glm10,glm9,glm0)
```

	Resid. Df <dbl>	Resid. Dev <dbl>	Df <dbl>	Deviance <dbl>
1	24954	15715.37	NA	NA
2	24976	16365.23	-22	-649.86901
3	24980	16387.06	-4	-21.82443
4	24999	27549.38	-19	-11162.32633

4 rows

```
pchisq( 21.824, 4, lower.tail = FALSE)
```

```
[1] 0.0002172518
```

```
pchisq( 145.728, 22, lower.tail = FALSE)
```

```
[1] 3.047793e-20
```

```
pchisq( 11162.3, 19, lower.tail = FALSE)
```

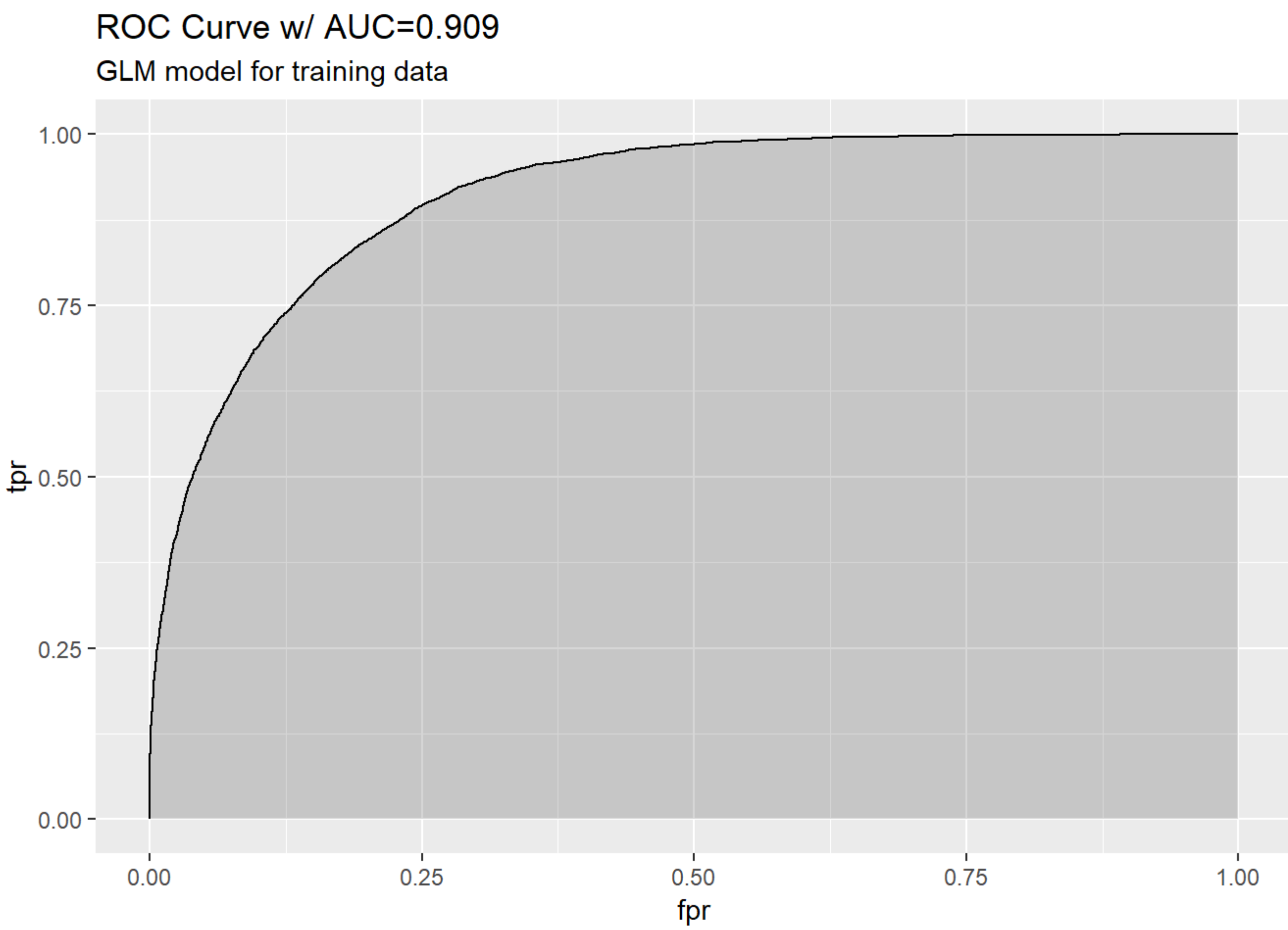
```
[1] 0
```

The anoav result suggest that the model including all the variables except [redacted] is the best fitted model.

ROC for the best model in best subset method

```
prob <- predict(glm12, data = train, type="response")
pred <- prediction(prob, train$income)
# rest of this doesn't need much adjustment except for titles
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
auc <- performance(pred, measure="auc")
auc <- round(auc@y.values[[1]],3)
roc.data <- data.frame(fpr=unlist(perf@x.values),
                       tpr=unlist(perf@y.values),
                       model="GLM")
ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
    geom_ribbon(alpha=0.2) +
```

```
    geom_line(aes(y=tpr)) +
    labs(title = paste0("ROC Curve w/ AUC=", auc),
         subtitle = "GLM model for training data")
```



Based on the C statistic (AUC = 0.909) this would rank somewhere near the high end of a pretty good predictive model by the ROC curve standard.

Check the Predictability of best subset model

```
glm.probs = predict(glm12, test, type ="response")
glm.pred = rep(0, length(glm.probs))
glm.pred[glm.probs >0.5] <- 1
table(glm.pred, test$income)
```

```

glm.pred    0    1
       0 5310  773
       1  408 1070
```

```
mean(glm.pred!= test$income)
```

```
[1] 0.1561963
```

For the best subset model, the test error rate is 15.59%, which is pretty good.

Forward and Backward Stepwise Selection

Forward selection

```
with(train,
     step(glm(income ~ 1,family = binomial),
          scope=(~ age+workclass+education+education.num+marital.status
                 +occupation+relationship+race+sex+capital.gain+
                   capital.loss+hours.per.week+native.country),
     direction="forward"))
```

```
Start:  AIC=27551.38
income ~ 1

                 Df Deviance   AIC
+ relationship    5    21750 21762
+ marital.status  6    22015 22029
+ occupation     14    24357 24387
+ education.num   1    24434 24438
+ education       2    24919 24925
+ capital.gain    1    24926 24930
+ hours.per.week  1    26114 26118
+ age             1    26180 26184
+ sex             1    26298 26302
+ workclass       8    26825 26843
+ capital.loss    1    27039 27043
+ race            4    27254 27264
+ native.country  1    27515 27519
<none>                 27549 27551

Step:  AIC=21761.98
income ~ relationship

                 Df Deviance   AIC
+ education.num   1    19089 19103
+ education       2    19442 19458
+ occupation     14    19429 19469
+ capital.gain    1    19679 19693
+ hours.per.week  1    21160 21174
+ workclass       8    21371 21399
+ capital.loss    1    21433 21447
+ age             1    21549 21563
+ marital.status  6    21617 21641
+ sex             1    21645 21659
+ race            4    21662 21682
+ native.country  1    21705 21719
<none>                 21750 21762

Step:  AIC=19103.37
```

```
income ~ relationship + education.num

                 Df Deviance   AIC
+ capital.gain    1    17482 17498
+ occupation     14    18372 18414
+ hours.per.week  1    18691 18707
+ age             1    18823 18839
+ workclass       8    18864 18894
+ capital.loss    1    18885 18901
+ marital.status  6    18911 18937
+ sex             1    18968 18984
+ native.country  1    19056 19072
+ race            4    19051 19073
+ education       2    19079 19097
<none>                 19089 19103

Step:  AIC=17497.57
income ~ relationship + education.num + capital.gain

                 Df Deviance   AIC
+ occupation     14    16825 16869
+ hours.per.week  1    17142 17160
+ capital.loss    1    17172 17190
+ workclass       8    17283 17315
+ age             1    17298 17316
+ marital.status  6    17328 17356
+ sex             1    17378 17396
+ race            4    17447 17471
+ native.country  1    17456 17474
+ education       2    17471 17491
<none>                 17482 17498

Step:  AIC=16869.18
income ~ relationship + education.num + capital.gain + occupation

                 Df Deviance   AIC
+ capital.loss    1    16546 16592
+ hours.per.week  1    16553 16599
+ age             1    16625 16671
+ marital.status  6    16678 16734
+ sex             1    16720 16766
+ workclass       8    16721 16781
+ race            4    16803 16855
+ native.country  1    16811 16857
<none>                 16825 16869
+ education       2    16824 16872

Step:  AIC=16592.33
income ~ relationship + education.num + capital.gain + occupation + 
    capital.loss

                 Df Deviance   AIC
+ hours.per.week  1    16285 16333
+ age             1    16363 16411
```

```
+ marital.status  6    16401 16459
+ sex             1    16444 16492
+ workclass       8    16448 16510
+ race            4    16526 16580
+ native.country  1    16533 16581
<none>                 16546 16592
+ education       2    16545 16595

Step:  AIC=16332.95
income ~ relationship + education.num + capital.gain + occupation + 
    capital.loss + hours.per.week

                 Df Deviance   AIC
+ age             1    16027 16077
+ marital.status  6    16144 16204
+ sex             1    16197 16247
+ workclass       8    16200 16264
+ race            4    16266 16322
+ native.country  1    16272 16322
<none>                 16285 16333
+ education       2    16284 16336

Step:  AIC=16077.06
income ~ relationship + education.num + capital.gain + occupation + 
    capital.loss + hours.per.week + age

                 Df Deviance   AIC
+ sex             1    15931 15983
+ marital.status  6    15930 15992
+ workclass       8    15930 15996
+ native.country  1    16017 16069
+ race            4    16013 16071
<none>                 16027 16077
+ education       2    16026 16080

Step:  AIC=15982.73
income ~ relationship + education.num + capital.gain + occupation + 
    capital.loss + hours.per.week + age + sex

                 Df Deviance   AIC
+ marital.status  6    15830 15894
+ workclass       8    15833 15901
+ native.country  1    15919 15973
+ race            4    15917 15977
<none>                 15931 15983
+ education       2    15929 15985

Step:  AIC=15894.37
income ~ relationship + education.num + capital.gain + occupation + 
    capital.loss + hours.per.week + age + sex + marital.status

                 Df Deviance   AIC
+ workclass       8    15733 15813
+ native.country  1    15818 15884
```

```
+ race            4    15816 15888
<none>                 15830 15894
+ education       2    15829 15897

Step:  AIC=15812.55
income ~ relationship + education.num + capital.gain + occupation +
    capital.loss + hours.per.week + age + sex + marital.status +
    workclass

                 Df Deviance   AIC
+ race            4    15716 15804
<none>                 15733 15813
+ education       2    15731 15815
+ native.country  1    16226 16308

Step:  AIC=15804.36
income ~ relationship + education.num + capital.gain + occupation +
    capital.loss + hours.per.week + age + sex + marital.status +
    workclass + race

                 Df Deviance   AIC
+ native.country  1    15707 15797
<none>                 15716 15804
+ education       2    15715 15807

Step:  AIC=15797.15
income ~ relationship + education.num + capital.gain + occupation +
    capital.loss + hours.per.week + age + sex + marital.status +
    workclass + race + native.country

            Df Deviance   AIC
<none>            15707 15797
+ education  2    15704 15798
```

```

Call:  glm(formula = income ~ relationship + education.num + capital.gain +
    occupation + capital.loss + hours.per.week + age + sex +
    marital.status + workclass + race + native.country, family = binomial)

Coefficients:
                         (Intercept)             relationship Not-in-family           relations
hip Other-relative                relationship Own-child                relationship Unmarried

                          -1.056e+01                             6.644e-01
        -1.745e-01                            -5.776e-01                             5.054e-01

                   relationship Wife                         education.num
      capital.gain                occupation Adm-clerical               occupation Armed-Forces

                           1.416e+00                             2.813e-01
         3.255e-04                            -6.105e+12                            -6.105e+12

             occupation Craft-repair            occupation Exec-managerial             occupati
```

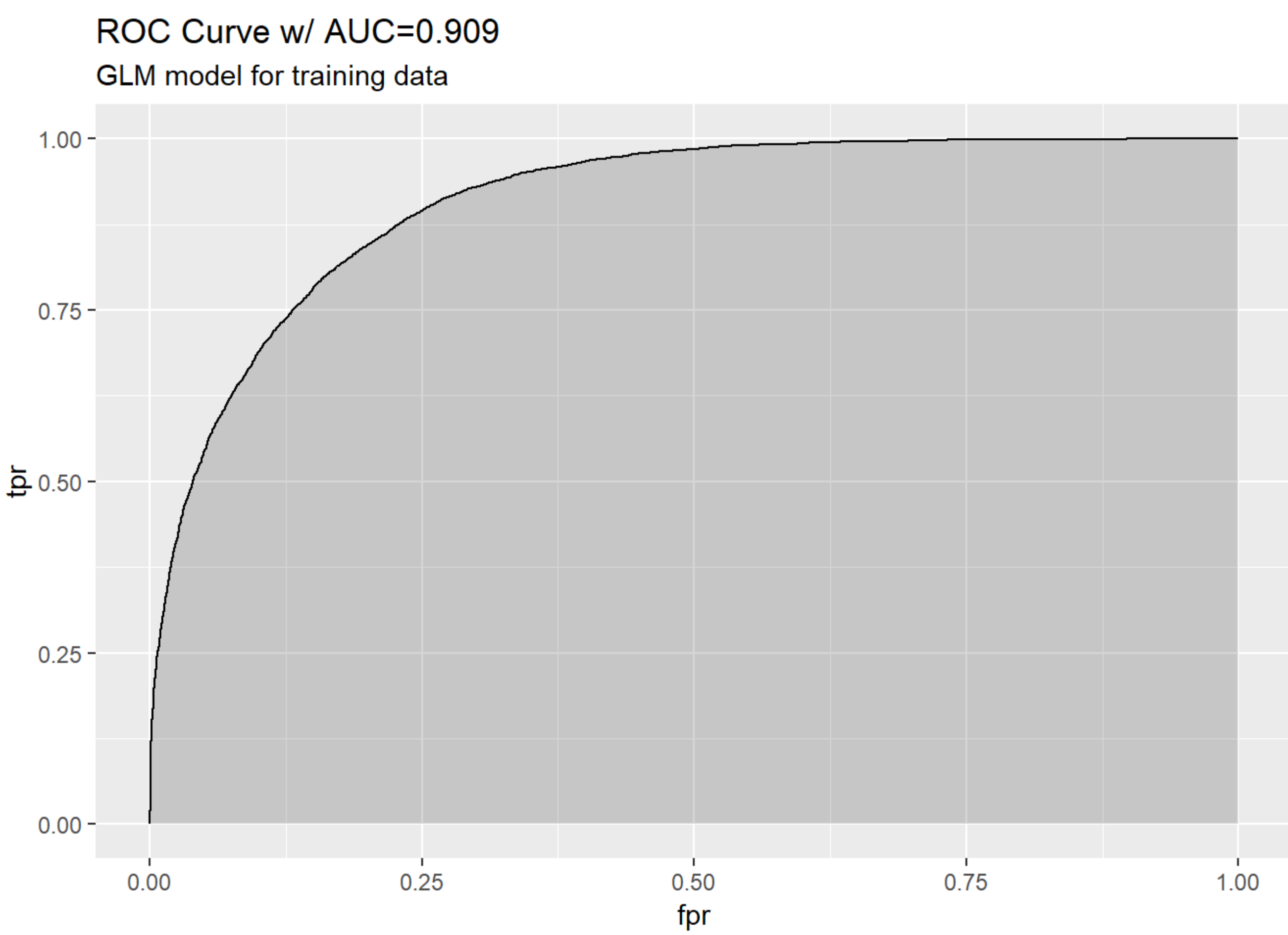
```
on Farming-fishing          occupation Handlers-cleaners          occupation Machine-op-inspct

                          -6.105e+12                            -6.105e+12
      -6.105e+12                            -6.105e+12                            -6.105e+12

            occupation Other-service            occupation Priv-house-serv              occupat
ion Prof-specialty            occupation Protective-serv                       occupation Sales

                          -6.105e+12                            -6.105e+12
      -6.105e+12                            -6.105e+12                            -6.105e+12

             occupation Tech-support           occupation Transport-moving
      capital.loss                       hours.per.week                                     age

                          -6.105e+12                            -6.105e+12
       6.473e-04                             3.157e-02                             2.732e-02

                            sex Male      marital.status Married-AF-spouse     marital.status
Married-civ-spouse  marital.status Married-spouse-absent         marital.status Never-married

                           9.002e-01                             2.997e+00
       2.283e+00                            -4.045e-02                            -5.259e-01

            marital.status Separated                marital.status Widowed                  wor
kclass Federal-gov                   workclass Local-gov               workclass Never-worked

                          -4.535e-02                            -3.656e-03
       6.105e+12                             6.105e+12                            -1.698e+01

                   workclass Private                workclass Self-emp-inc            workclas
s Self-emp-not-inc                   workclass State-gov                workclass Without-pay

                           6.105e+12                             6.105e+12
       6.105e+12                             6.105e+12                             6.105e+12

             race Asian-Pac-Islander                            race Black
      race Other                            race White                 native.countryNon-US

                           6.853e-01                             5.423e-01
       4.476e-03                             6.508e-01                            -2.569e-01


Degrees of Freedom: 24999 Total (i.e. Null);  24955 Residual
Null Deviance:      27550
Residual Deviance: 15710    AIC: 15800
```

Forward selection includes 11 variables.

```
fwd <- glm(formula = income ~ relationship + education.num + capital.gain +
    occupation + capital.loss + hours.per.week + age + sex +
    marital.status + workclass + native.country, family = binomial,data = train)
```

ROC for forward selection model

```
prob <- predict(fwd, data = train, type="response")
pred <- prediction(prob, train$income)

perf <- performance(pred, measure = "tpr", x.measure = "fpr")
auc <- performance(pred, measure="auc")
auc <- round(auc@y.values[[1]],3)
roc.data <- data.frame(fpr=unlist(perf@x.values),
                       tpr=unlist(perf@y.values),
                       model="GLM")
ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
    geom_ribbon(alpha=0.2) +
    geom_line(aes(y=tpr)) +
    labs(title = paste0("ROC Curve w/ AUC=", auc),
         subtitle = "GLM model for training data")
```



Check Predictability

```
fwd.probs = predict(fwd, test, type ="response")
fwd.pred = rep(0, length(fwd.probs))
fwd.pred[fwd.probs >0.5] <- 1
table(fwd.pred, test$income)
```

```
fwd.pred    0    1
       0 5309  791
       1  409 1052
```

```
mean(fwd.pred!= test$income)
```

```
[1] 0.1587092
```

For the model selected by forward selection method, the test error rate is 15.83%, which is pretty good.

Backward selection

```
with(train,
     step(glm(income ~ age+workclass+education+education.num+marital.status
                 +occupation+relationship+race+sex+capital.gain+
                   capital.loss+hours.per.week+native.country,family = binomial),
          direction="backward"))
```

```
Start:  AIC=15795.68
income ~ age + workclass + education + education.num + marital.status +
    occupation + relationship + race + sex + capital.gain + capital.loss +
    hours.per.week + native.country

                 Df Deviance   AIC
- education       2    15705 15793
<none>                 15704 15796
- race            4    15717 15801
- native.country  1    15715 15805
- workclass       7    15804 15882
- marital.status  6    15806 15886
- sex             1    15803 15893
- education.num   1    15886 15976
- age             1    15921 16011
- relationship    5    15929 16011
- capital.loss    1    15945 16035
- hours.per.week  1    16003 16093
- occupation     13    16158 16224
- capital.gain    1    17165 17255

Step:  AIC=15793.45
income ~ age + workclass + education.num + marital.status + occupation +
    relationship + race + sex + capital.gain + capital.loss +
    hours.per.week + native.country

                 Df Deviance   AIC
<none>                 15705 15793
- race            4    15718 15798
- native.country  1    15716 15802
- workclass       7    15806 15880
```

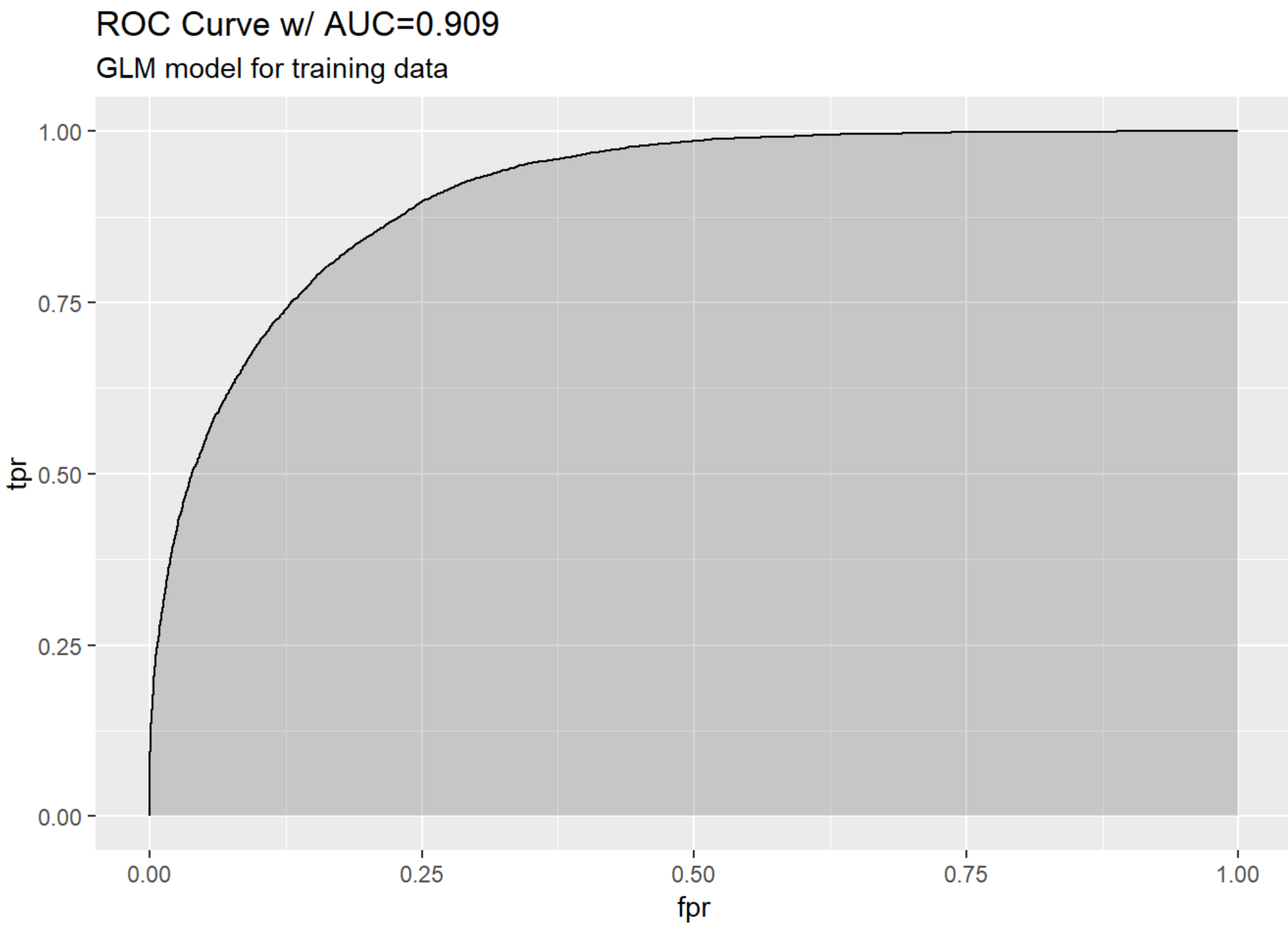
```
- marital.status  6    15808 15884
- sex             1    15805 15891
- relationship    5    15931 16009
- age             1    15925 16011
- capital.loss    1    15947 16033
- hours.per.week  1    16004 16090
- occupation     13    16172 16234
- education.num   1    16509 16595
- capital.gain    1    17167 17253
```

```

Call:  glm(formula = income ~ age + workclass + education.num + marital.status +
    occupation + relationship + race + sex + capital.gain + capital.loss +
    hours.per.week + native.country, family = binomial)

Coefficients:
                         (Intercept)                                  age                  wor
kclass Federal-gov                   workclass Local-gov                workclass Never-worked

                          -1.056e+01                            2.736e-02
         1.032e+00                            2.927e-01                            -9.550e+00

                   workclass Private                workclass Self-emp-inc            workclas
s Self-emp-not-inc                   workclass State-gov                 workclass Without-pay

                           5.469e-01                            6.676e-01
         4.421e-02                            1.888e-01                            -1.153e+01

                       education.num      marital.status Married-AF-spouse     marital.status
Married-civ-spouse  marital.status Married-spouse-absent          marital.status Never-married

                           2.813e-01                            2.987e+00
         2.282e+00                           -4.260e-02                            -5.250e-01

           marital.status Separated                marital.status Widowed                occup
ation Adm-clerical                occupation Armed-Forces               occupation Craft-repair

                          -4.399e-02                           -3.796e-03
         7.671e-02                           -6.984e-01                             1.464e-01

          occupation Exec-managerial            occupation Farming-fishing           occupation
 Handlers-cleaners           occupation Machine-op-inspct              occupation Other-service

                           8.787e-01                           -8.196e-01
        -5.000e-01                           -2.893e-01                            -8.330e-01

          occupation Priv-house-serv             occupation Prof-specialty            occupati
on Protective-serv                        occupation Sales               occupation Tech-support

                          -3.574e+00                            6.424e-01
         6.715e-01                            3.432e-01                             7.450e-01

         occupation Transport-moving            relationship Not-in-family           relations
```

```
hip Other-relative                  relationship Own-child                relationship Unmarried

                              NA                            6.671e-01
        -1.651e-01                              -5.749e-01                              5.132e-01

               relationship Wife                race Asian-Pac-Islander
        race Black                              race Other                              race White

                        1.411e+00                       6.850e-01
         5.429e-01                               3.622e-03                              6.508e-01

                         sex Male                          capital.gain
      capital.loss                          hours.per.week                  native.countryNon-US

                        9.016e-01                       3.257e-04
         6.482e-04                               3.159e-02                             -2.560e-01


Degrees of Freedom: 24999 Total (i.e. Null);  24956 Residual
Null Deviance:      27550
Residual Deviance: 15710    AIC: 15790
```

```
bwd<- glm(formula = income ~ age + workclass + education.num + marital.status +
    occupation + relationship + race + sex + capital.gain + capital.loss +
    hours.per.week + native.country, family = binomial(link = logit),data = train)
```

```
prob <- predict(bwd, data = train, type="response")
pred <- prediction(prob, train$income)

perf <- performance(pred, measure = "tpr", x.measure = "fpr")
auc <- performance(pred, measure="auc")
auc <- round(auc@y.values[[1]],3)
roc.data <- data.frame(fpr=unlist(perf@x.values),
                       tpr=unlist(perf@y.values),
                       model="GLM")
ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
    geom_ribbon(alpha=0.2) +
    geom_line(aes(y=tpr)) +
    labs(title = paste0("ROC Curve w/ AUC=", auc),
         subtitle = "GLM model for training data")
```



The backward method returns thefinal model, which includes 12 variables.

Check Predictability

```
bwd.probs = predict(bwd, test, type ="response")
bwd.pred = rep(0, length(bwd.probs))
bwd.pred[bwd.probs >0.5] <- 1
table(bwd.pred, test$income)
```

```

bwd.pred    0    1
       0 5310  776
       1  408 1067
```

```
mean(bwd.pred!= test$income)
```

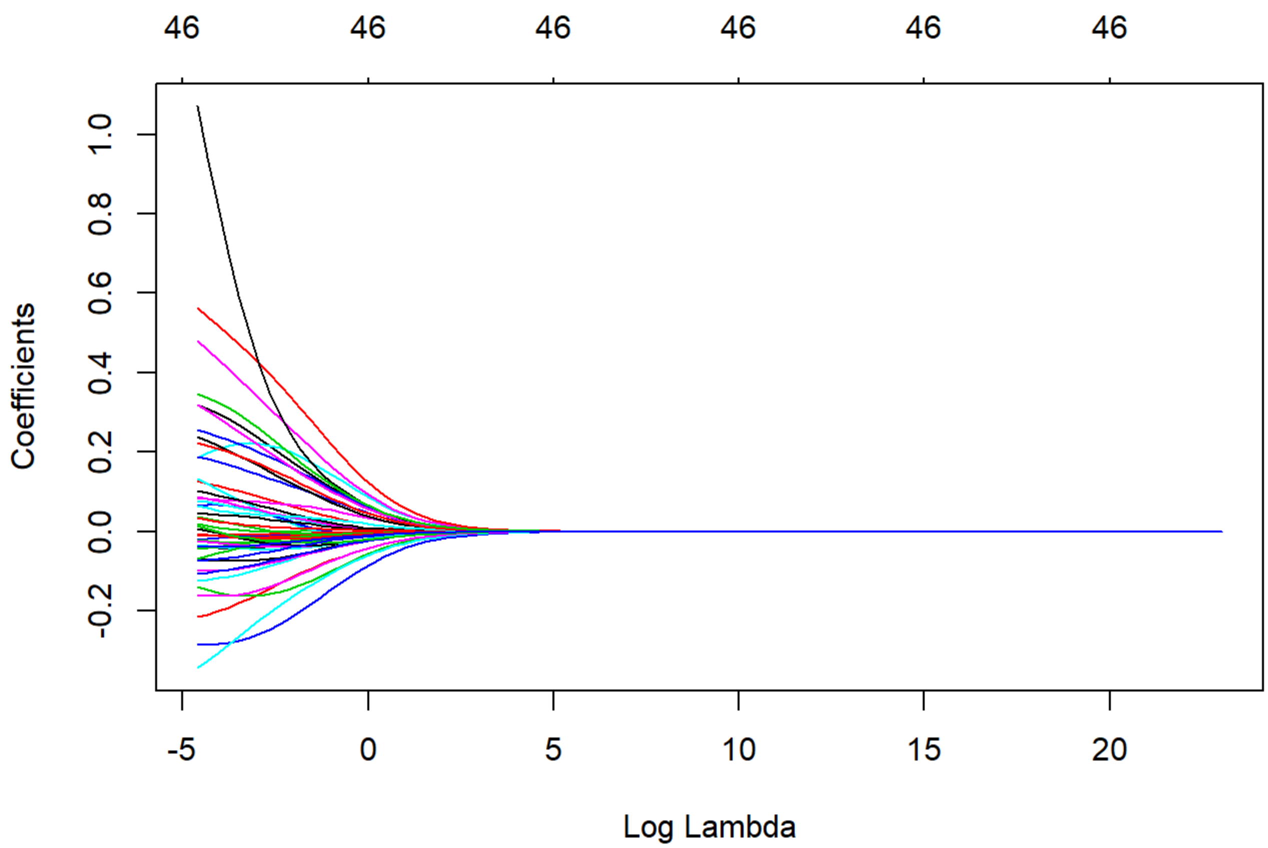
```
[1] 0.156593
```

For the model selected by forward selection method, the test error rate is 15.66%, which is pretty good.

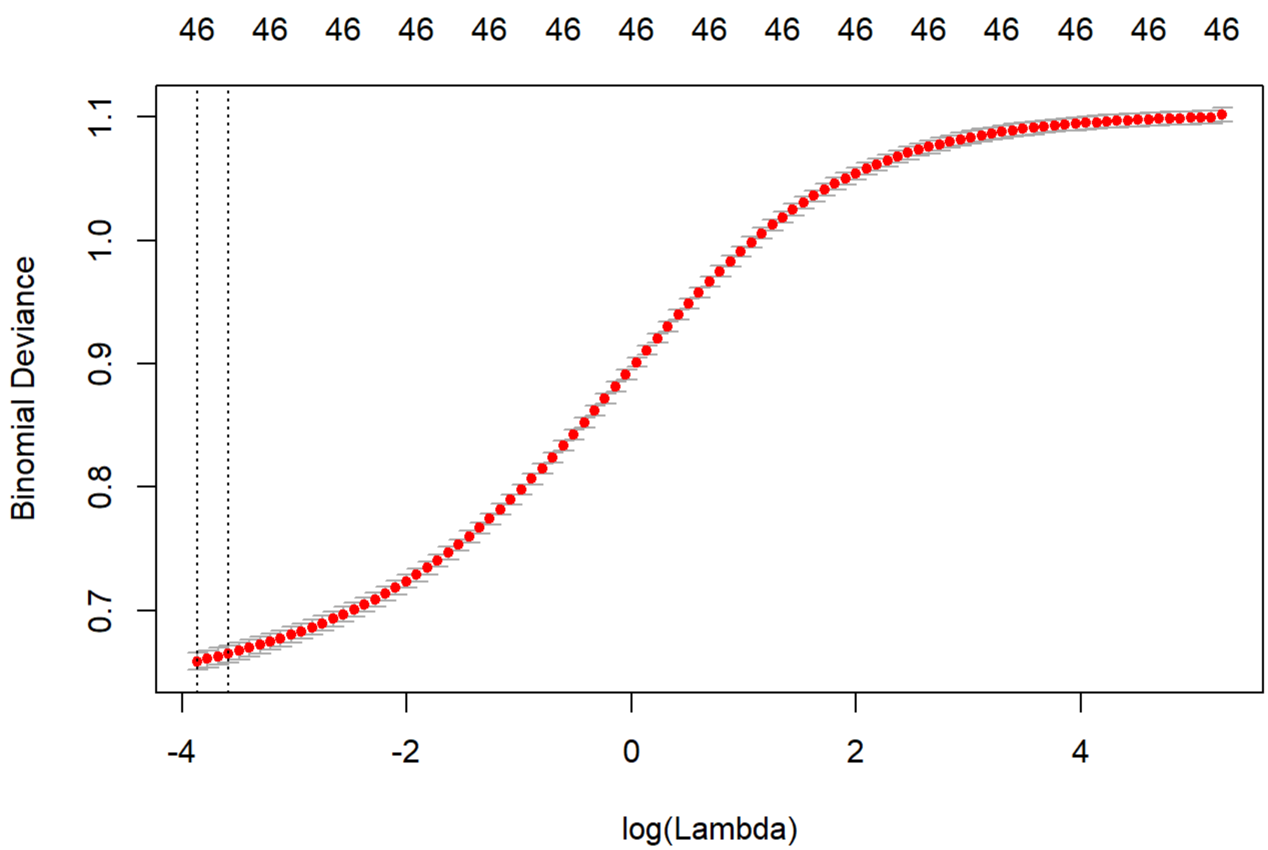
Ridge Regression

```r
x <- model.matrix(income ~ .-fnlwgt, data = train)[,-1]
y <- train$income
x <- scale(x)
grid<- 10^seq(10,-2,length=100)
```

```r
library(glmnet)
ridge <- glmnet(x, y, alpha = 0, lambda = grid, family = "binomial")
plot(ridge, xvar="lambda")
```




```r
cv.ridge <-cv.glmnet(x,y,alpha=0, family = "binomial")
plot(cv.ridge)
```




```
bestlam.ridge <- cv.ridge$lambda.min
bestlam.ridge
```

```
[1] 0.02099622
```

```
predict(ridge, s=bestlam.ridge, type = "coefficients")
```

```
47 x 1 sparse Matrix of class "dgCMatrix"
                                           1
(Intercept)                     -1.855759151
age                              0.289434628
workclass Federal-gov            0.111329436
workclass Local-gov              0.008521076
workclass Never-worked          -0.014484927
workclass Private                0.097878553
workclass Self-emp-inc           0.079155547
workclass Self-emp-not-inc      -0.041950237
workclass State-gov             -0.011920739
workclass Without-pay           -0.035053995
educationsome college            0.066818893
educationBachelors and above     0.213565708
education.num                    0.418365476
marital.status Married-AF-spouse 0.041965444
marital.status Married-civ-spouse 0.504075976
```

```
marital.status Married-spouse-absent -0.026491356
marital.status Never-married         -0.281320602
marital.status Separated             -0.039902121
marital.status Widowed               -0.029801982
occupation Adm-clerical              -0.007349382
occupation Armed-Forces              -0.009210186
occupation Craft-repair               0.024678413
occupation Exec-managerial            0.233686910
occupation Farming-fishing           -0.114311636
occupation Handlers-cleaners         -0.096521078
occupation Machine-op-inspct         -0.073527425
occupation Other-service             -0.194230746
occupation Priv-house-serv           -0.049424658
occupation Prof-specialty             0.170530501
occupation Protective-serv            0.067272749
occupation Sales                      0.073601671
occupation Tech-support               0.087656927
occupation Transport-moving          -0.012222172
relationship Not-in-family           -0.156219018
relationship Other-relative          -0.096520921
relationship Own-child               -0.291946641
relationship Unmarried               -0.161962685
relationship Wife                     0.209129204
race Asian-Pac-Islander               0.023210063
race Black                           -0.004583987
race Other                           -0.038430048
race White                            0.050816876
sex Male                              0.276202535
capital.gain                          0.742737936
capital.loss                          0.203695906
hours.per.week                        0.316550631
native.countryNon-US                 -0.067555136
```

The smallest lambda is 0.021 using the cross-validation methods.

```
testx <- model.matrix(income~.-fnlwgt, data = test)
ridge.probs = predict(ridge,  s = bestlam.ridge, newx = testx)
ridge.pred = rep(0, length(ridge.probs))
ridge.pred[ridge.probs >0.5] <- 1
table(ridge.pred, test$income)
```

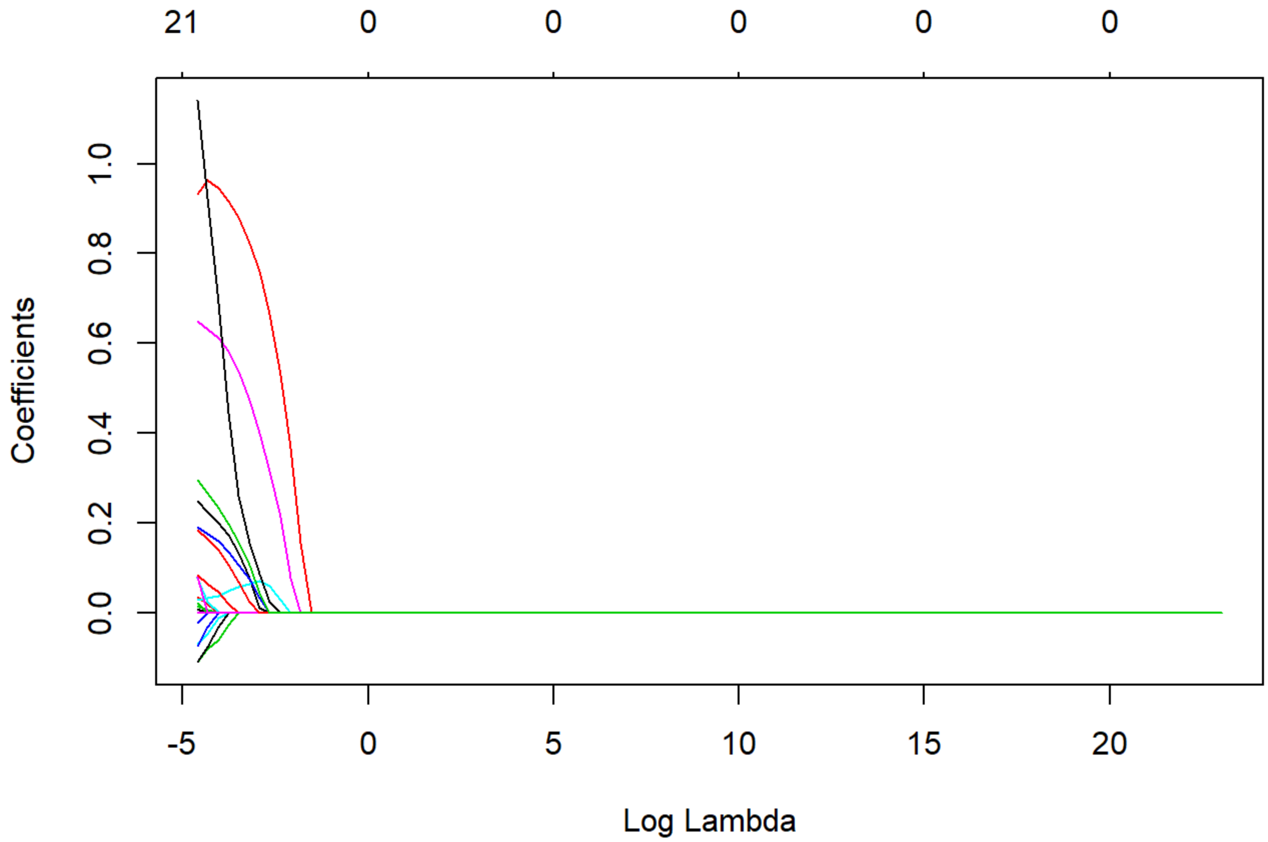
```


ridge.pred    0    1
         1 5718 1843
```

```
mean(ridge.pred!= test$income)
```

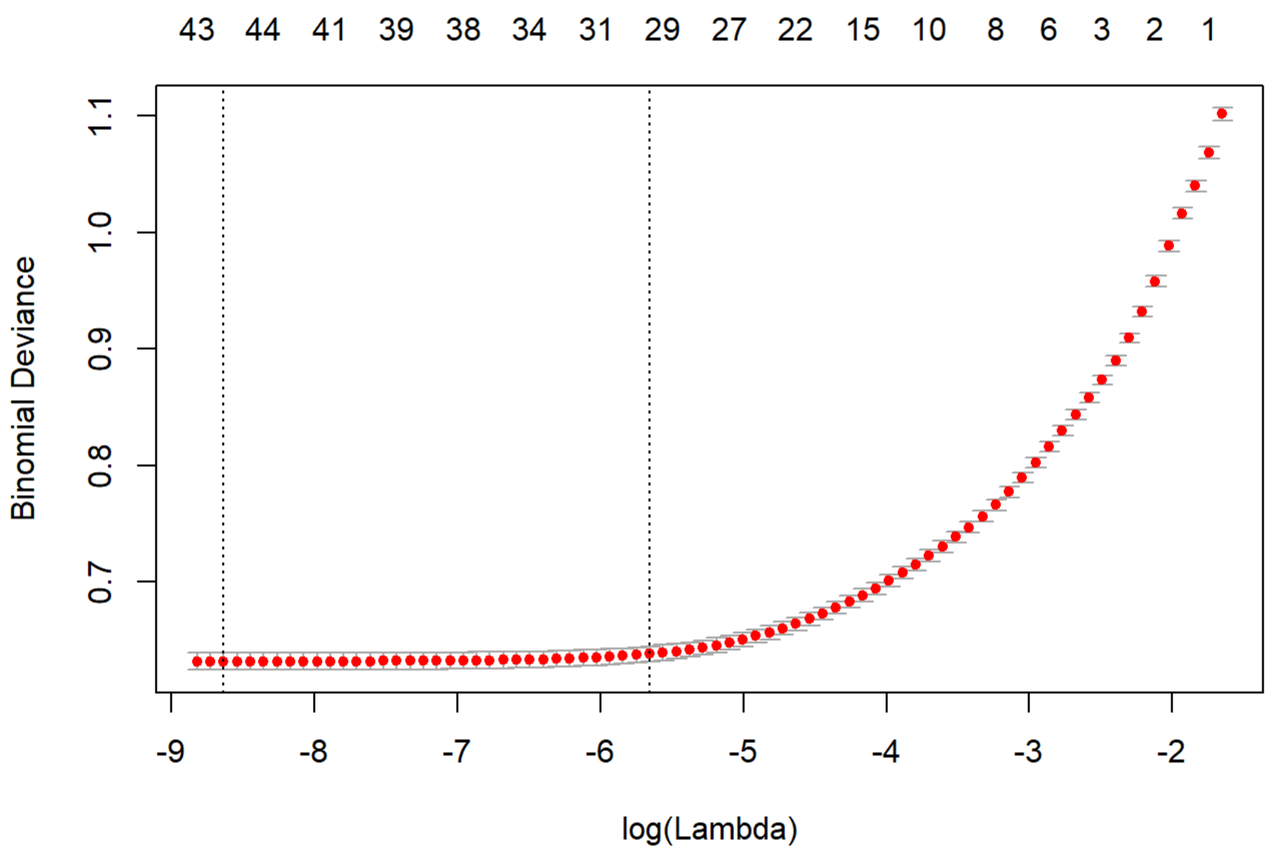
```
[1] 0.7562492
```

lasso Regression

```
lasso <- glmnet(x, y, alpha = 1, lambda = grid, family = "binomial")
plot(lasso, xvar="lambda")
```




```
cv.lasso <-cv.glmnet(x,y,alpha=1, family = "binomial")
plot(cv.lasso)
```




```
bestlam.lasso <- cv.lasso$lambda.min
bestlam.lasso
```

```
[1] 0.0001784161
```

The smallest lambda is 0.000235 using the cross-validation methods.

```
predict(lasso, s=bestlam.lasso, type = "coefficients")
```

```
47 x 1 sparse Matrix of class "dgCMatrix"
                                              1
(Intercept)                       -1.7168153871
age                                0.2491328870
workclass Federal-gov              0.0367936628
workclass Local-gov                .
workclass Never-worked             .
workclass Private                  .
workclass Self-emp-inc             0.0149392435
workclass Self-emp-not-inc        -0.0212565031
workclass State-gov                .
workclass Without-pay              .
educationsome college              .
educationBachelors and above       0.0285296551
education.num                      0.6491686280
```

```
marital.status Married-AF-spouse      0.0098392643
marital.status Married-civ-spouse     0.9329462992
marital.status Married-spouse-absent  .
marital.status Never-married         -0.1077032920
marital.status Separated              .
marital.status Widowed                .
occupation Adm-clerical               .
occupation Armed-Forces               .
occupation Craft-repair               .
occupation Exec-managerial            0.1912710386
occupation Farming-fishing           -0.0686105117
occupation Handlers-cleaners          .
occupation Machine-op-inspct         -0.0005407155
occupation Other-service             -0.1091274725
occupation Priv-house-serv            .
occupation Prof-specialty             0.0839401636
occupation Protective-serv            .
occupation Sales                      .
occupation Tech-support               0.0228834379
occupation Transport-moving           .
relationship Not-in-family            .
relationship Other-relative           .
relationship Own-child               -0.0732674411
relationship Unmarried                .
relationship Wife                     0.0770372563
race Asian-Pac-Islander               .
race Black                            .
race Other                            .
race White                            .
sex Male                              0.0799454041
capital.gain                          1.1405111960
capital.loss                          0.1847009949
hours.per.week                        0.2963372712
native.countryNon-US                  .
```

```
testx <- model.matrix(income~.-fnlwgt, data = test)
lasso.probs = predict(lasso,s = bestlam.ridge, newx = testx)
lasso.pred = rep(0, length(lasso.probs))
lasso.pred[lasso.probs >0.5] <- 1
table(lasso.pred, test$income)
```

```
lasso.pred    0    1
         1 5718 1843
```

```
mean(lasso.pred!= test$income)
```

```
[1] 0.7562492
```

The models given by both lasso and ridge have a higher error rate than the model generated before.