STOR767 - HW 3 Computing

Leo Li (PID: 730031954)

Problem: IQ and successes

Background: Measurement of Intelligence

Case Study: how intelligence relates to one's future successes?

Data needed: `IQ.Full.csv`

ASVAB (Armed Services Vocational Aptitude Battery) tests have been used as a screening test for those who want to join the army or other jobs.

Our data set `IQ.csv` is a subset of individuals from the 1979 National Longitudinal Study of Youth (NLSY79) survey who were re-interviewed in 2006. Information about family, personal demographic such as gender, race and education level, plus a set of ASVAB (Armed Services Vocational Aptitude Battery) test scores are available. It is STILL used as a screening test for those who want to join the army! ASVAB scores were 1981 and income was 2005.

Our goals:

- **Is IQ related to one's successes measured by Income?**
- **Is there evidence to show that Females are under-paid?**
- **What are the best possible prediction models to predict future income?**

The ASVAB has the following components:

- **Science, Arith (Arithmetic reasoning), Word (Word knowledge), Parag (Paragraph comprehension), Numer (Numerical operation), Coding (Coding speed), Auto (Automative and Shop information), Math (Math knowledge), Mechanic (Mechanic Comprehension) and Elec (Electronic information).**
- **AFQT (Armed Forces Qualifying Test) is a combination of Word, Parag, Math and Arith.**
- **Note: Service Branch requirement: Army 31, Navy 35, Marines 31, Air Force 36, and Coast Guard 45, (out of 100 which is the max!)**

The detailed variable definitions:

Personal Demographic Variables:

- **Race: 1 = Hispanic, 2 = Black, 3 = Not Hispanic or Black**
- **Gender: a factor with levels "female" and "male"**
- **Educ: years of education completed by 2006**

Household Environment:

- **Imagazine: a variable taking on the value 1 if anyone in the respondent's household regularly read magazines in 1979, otherwise 0**
- **Inewspaper: a variable taking on the value 1 if anyone in the respondent's household regularly read newspapers in 1979, otherwise 0**
- **Ilibrary: a variable taking on the value 1 if anyone in the respondent's household had a library card in 1979, otherwise 0**
- **MotherEd: mother's years of education**

- **FatherEd: father's years of education**

Variables Related to ASVAB test Scores in 1981 (Proxy of IQ's)

- **AFQT: percentile score on the AFQT intelligence test in 1981**
- **Coding: score on the Coding Speed test in 1981**
- **Auto: score on the Automotive and Shop test in 1981**
- **Mechanic: score on the Mechanic test in 1981**
- **Elec: score on the Electronics Information test in 1981**
- **Science: score on the General Science test in 1981**
- **Math: score on the Math test in 1981**
- **Arith: score on the Arithmetic Reasoning test in 1981**
- **Word: score on the Word Knowledge Test in 1981**
- **Parag: score on the Paragraph Comprehension test in 1981**
- **Numer: score on the Numerical Operations test in 1981**

Variable Related to Life Success in 2006

- **Income2005: total annual income from wages and salary in 2005. We will use a natural log transformation over the income.**

The following 10 questions are answered as 1: strongly agree, 2: agree, 3: disagree, 4: strongly disagree

- **Esteem 1: "I am a person of worth"????**
- **Esteem 2: "I have a number of good qualities"????**
- **Esteem 3: "I am inclined to feel like a failure"????**
- **Esteem 4: "I do things as well as others"????**
- **Esteem 5: "I do not have much to be proud of"????**
- **Esteem 6: "I take a positive attitude towards myself and others"????**
- **Esteem 7: "I am satisfied with myself"????**
- **Esteem 8: "I wish I could have more respect for myself"????**
- **Esteem 9: "I feel useless at times"????**
- **Esteem 10: "I think I am no good at all"????**

Note: we will not use the Esteem scores in this homework.

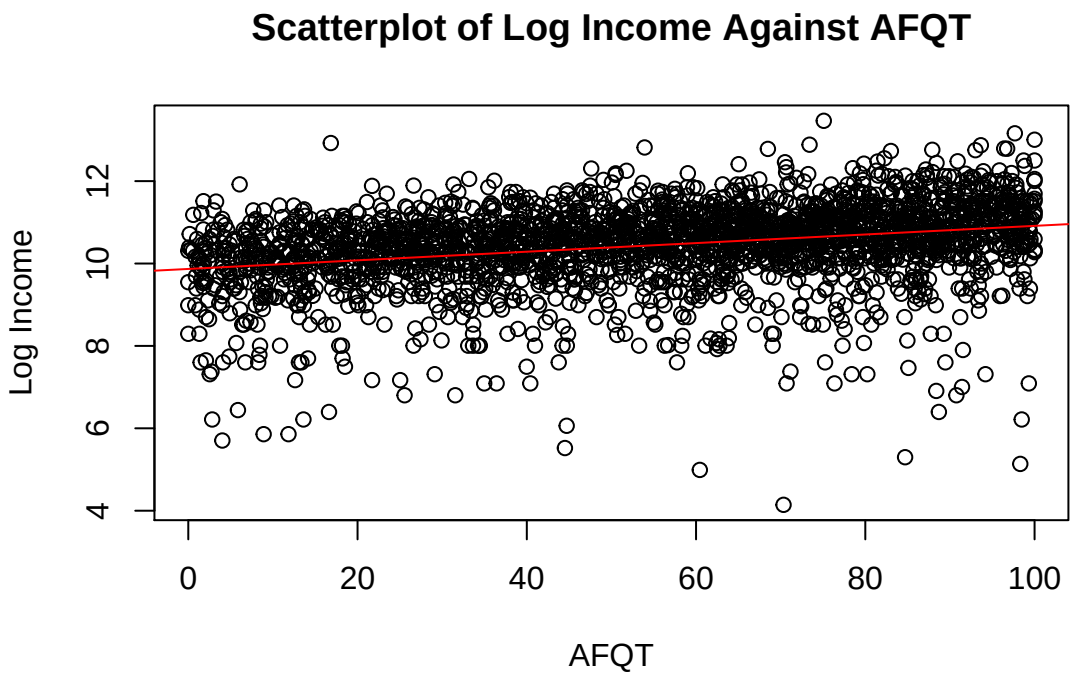
1. EDA

i. **Make a few informative tables or plots to see how intellegence might relate to the log income.**

In this question, we would like to make a few plots to see how intellegence might relate to the log income. Here, we make a scatterplot for each of the proxy of IQ and the log income to visualize their relationship, a linear regression line is also plotted to help visualize the trend.

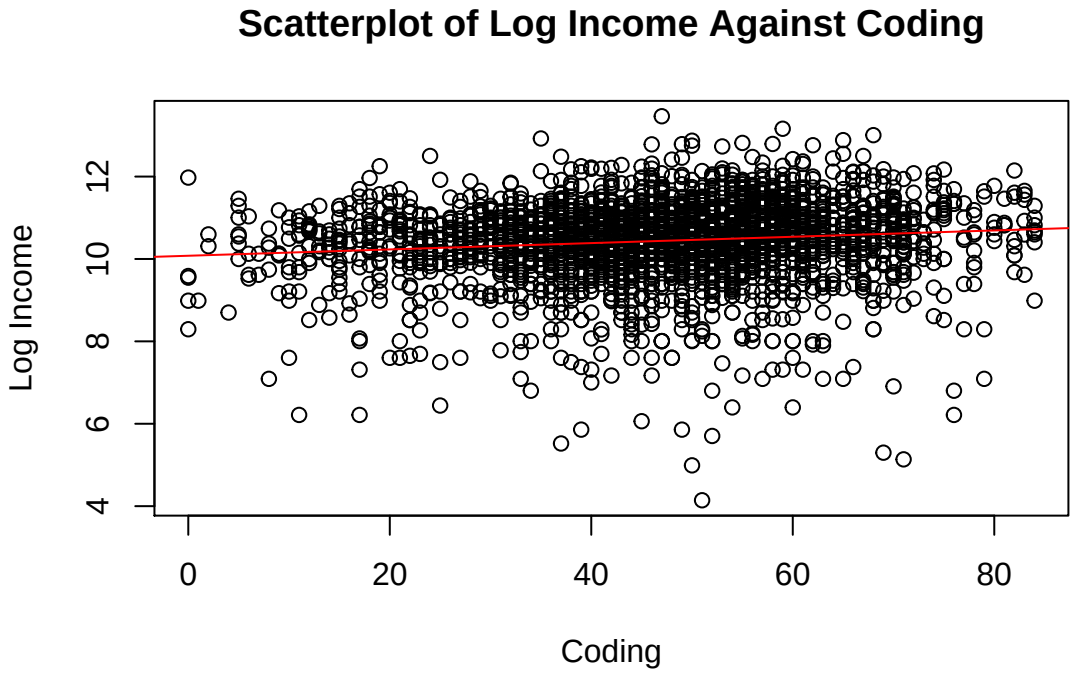
1. AFQT and log income

```
plot(iq$AFQT, iq$Income2005, main="Scatterplot of Log Income Against AFQT",
     xlab="AFQT", ylab="Log Income")
abline(lm(iq$Income2005~iq$AFQT), col="red")
lines(lowess(iq$Income2005,iq$AFQT), col="blue")
```



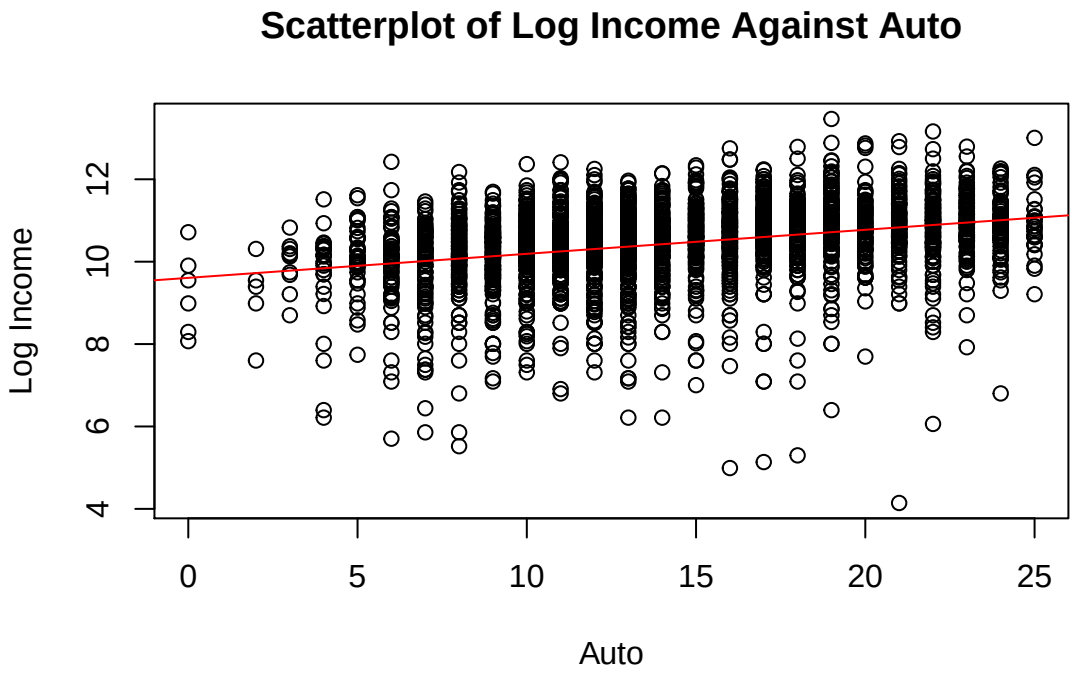
2. Coding and log income

```
plot(iq$Coding, iq$Income2005, main="Scatterplot of Log Income Against Coding",
    xlab="Coding",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Coding), col="red")
```



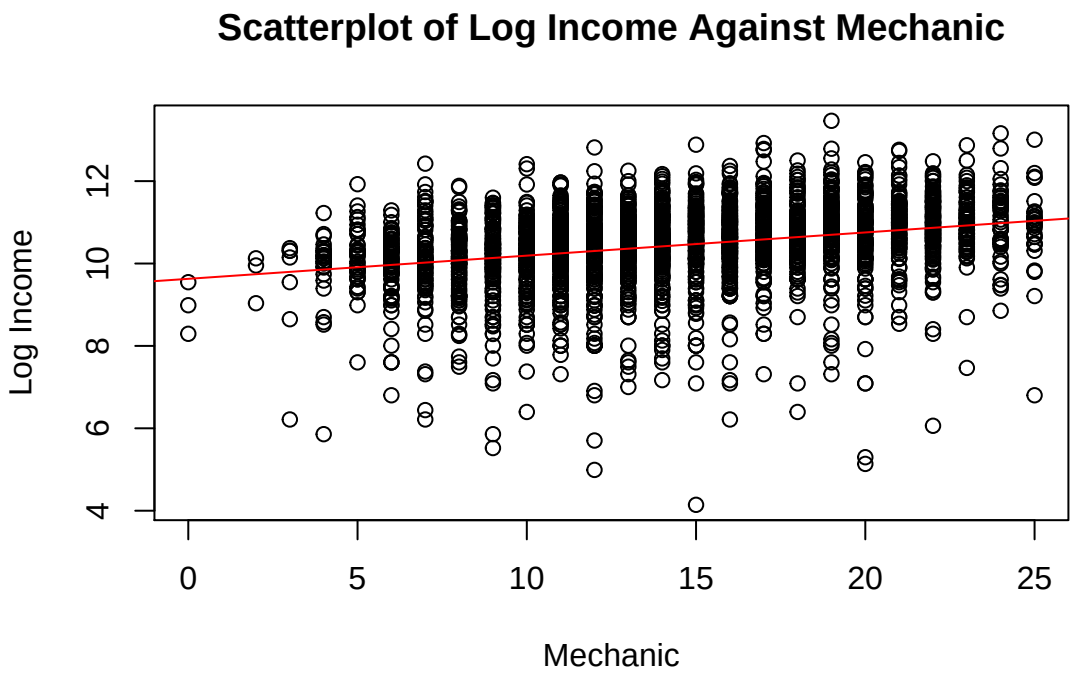
3. Auto and log income

```
plot(iq$Auto, iq$Income2005, main="Scatterplot of Log Income Against Auto",
    xlab="Auto",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Auto), col="red")
```



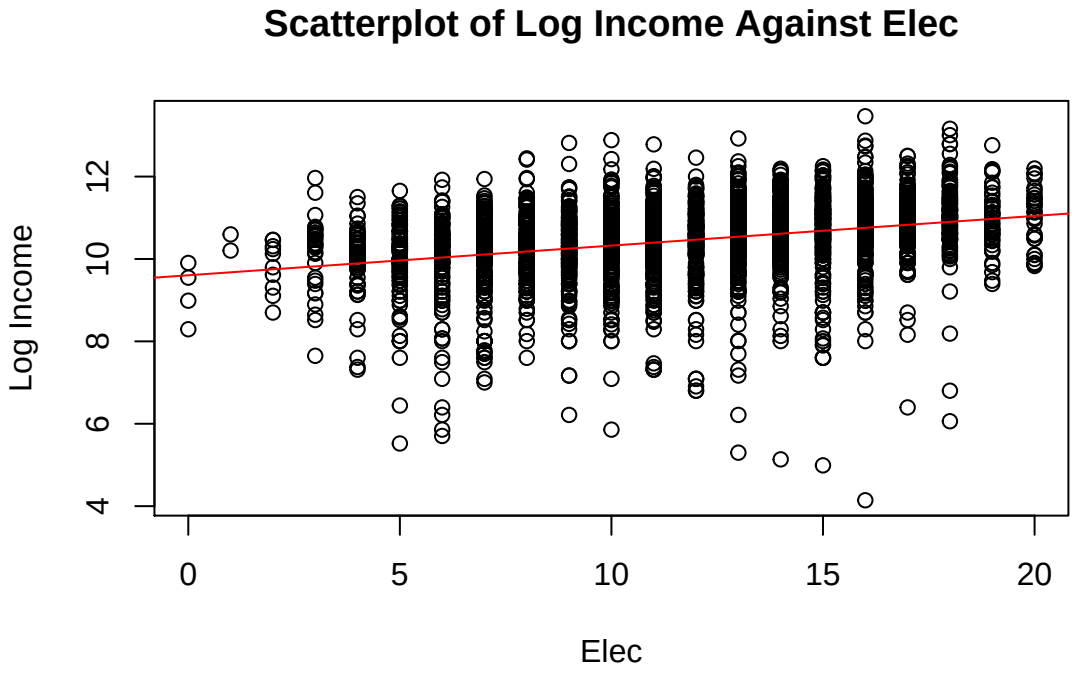
4. Mechanic and log income

```
plot(iq$Mechanic, iq$Income2005, main="Scatterplot of Log Income Against Mechanic",
    xlab="Mechanic",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Mechanic), col="red")
```



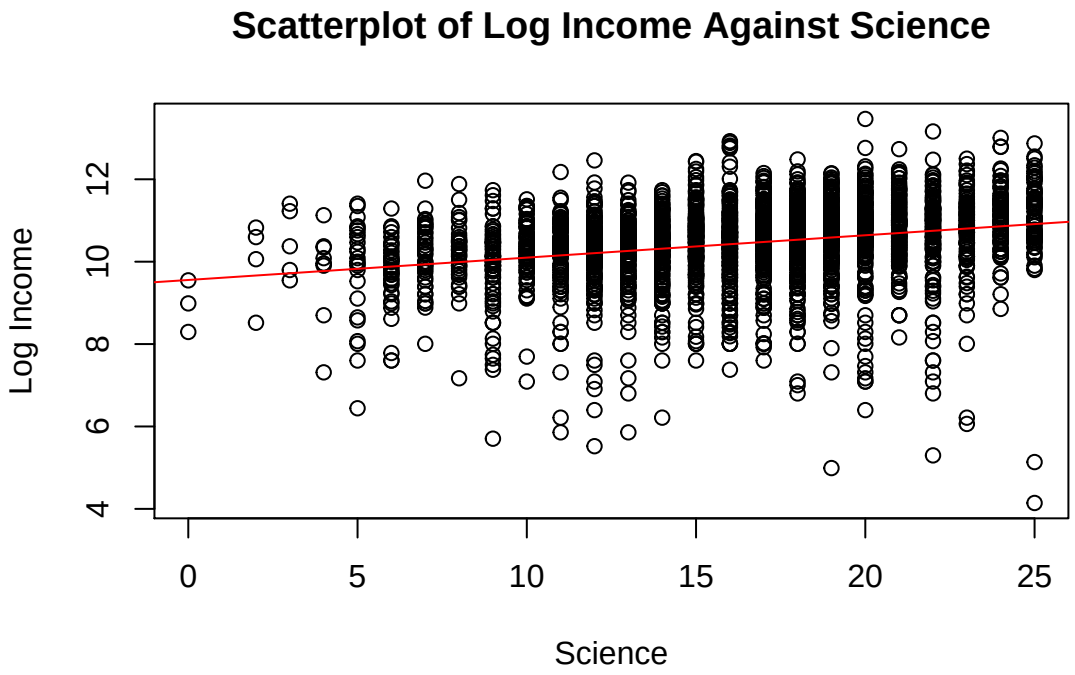
5. Elec and log income

```
plot(iq$Elec, iq$Income2005, main="Scatterplot of Log Income Against Elec",
    xlab="Elec",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Elec), col="red")
```



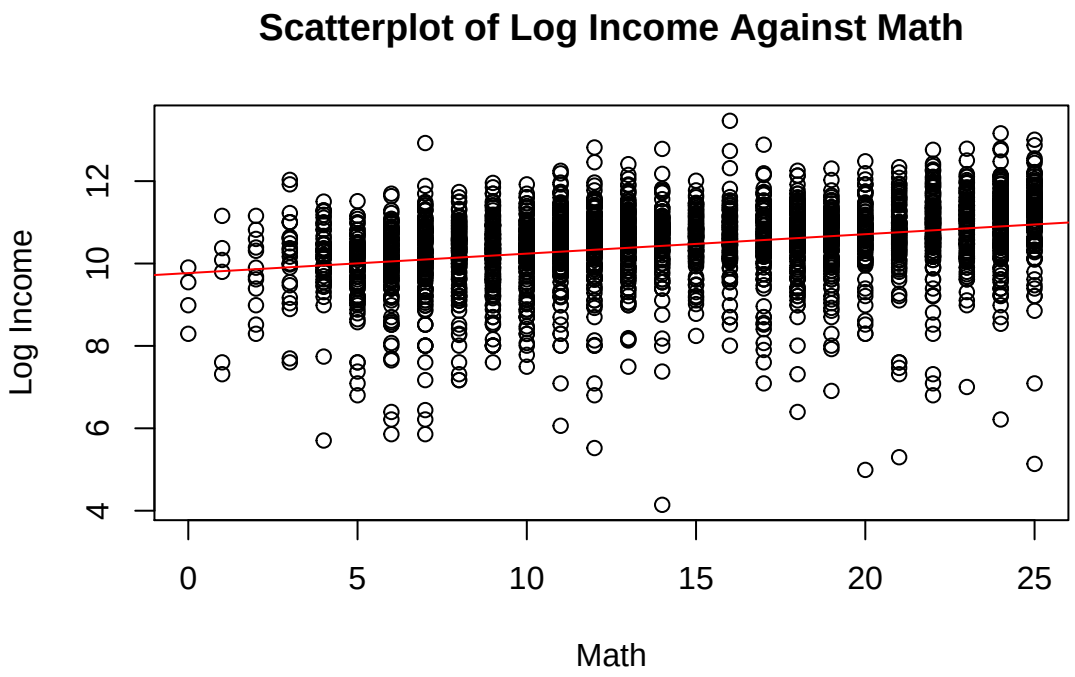
6. Science and log income

```
plot(iq$Science, iq$Income2005, main="Scatterplot of Log Income Against Science",
    xlab="Science",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Science), col="red")
```



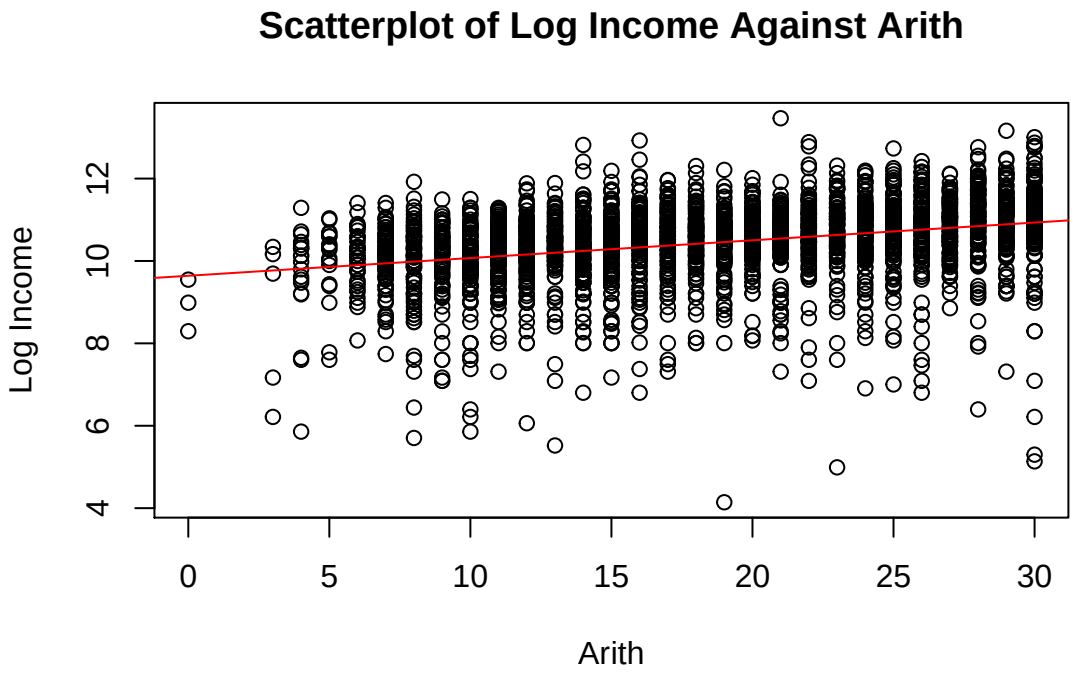
7. Math and log income

```
plot(iq$Math, iq$Income2005, main="Scatterplot of Log Income Against Math",
    xlab="Math",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Math), col="red")
```



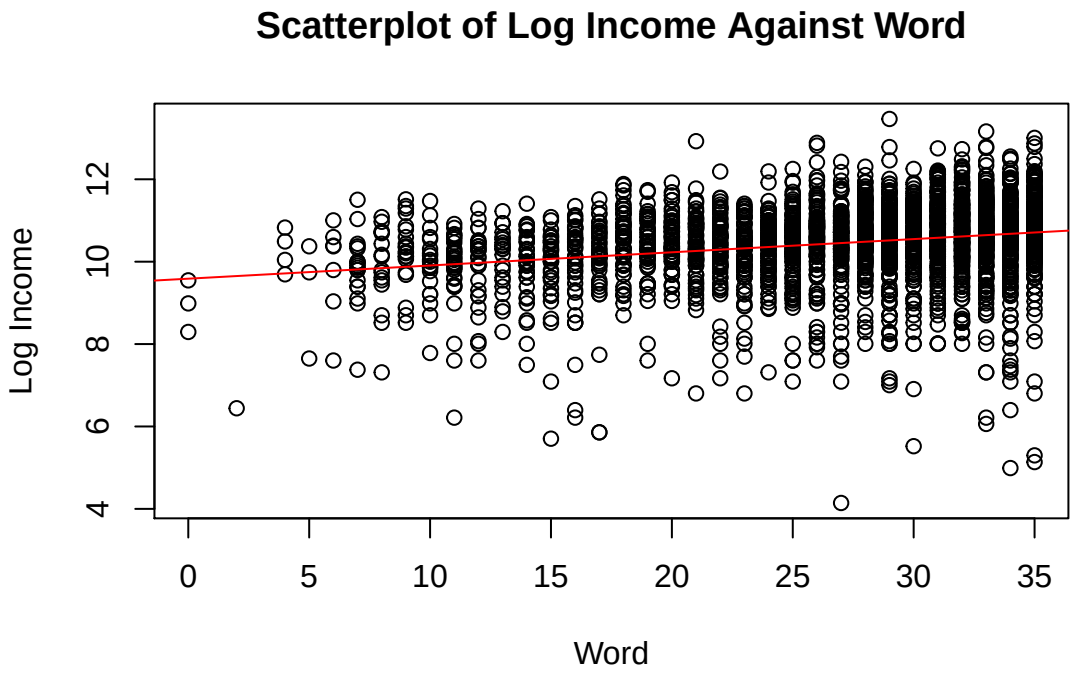
8. Arith and log income

```
plot(iq$Arith, iq$Income2005, main="Scatterplot of Log Income Against Arith",
    xlab="Arith",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Arith), col="red")
```



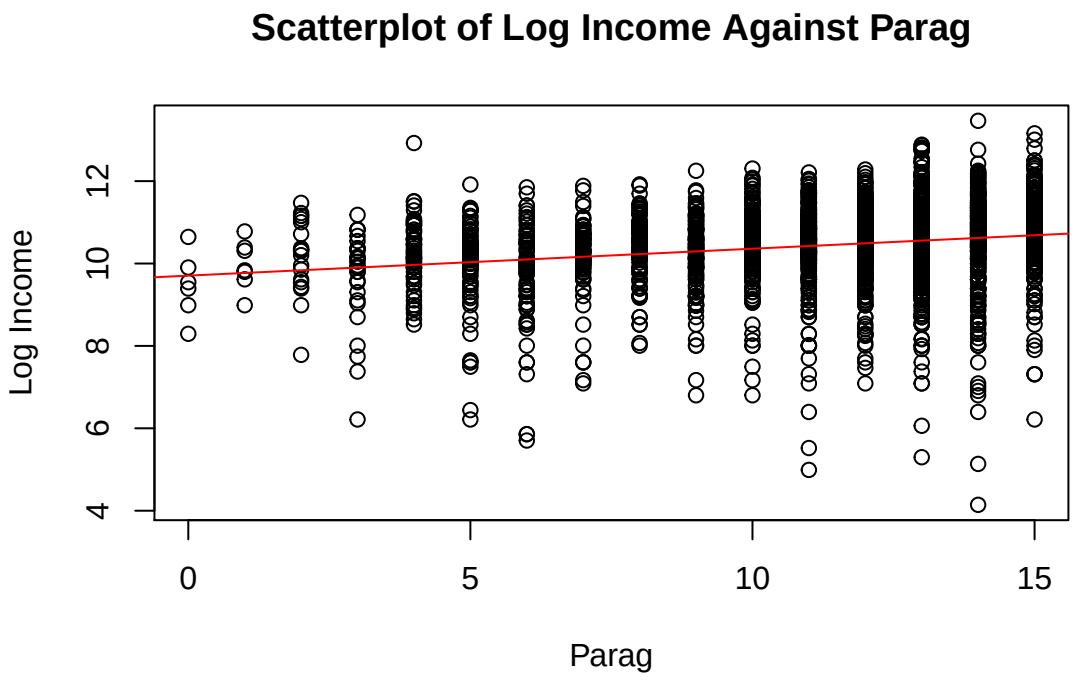
9. Word and log income

```
plot(iq$Word, iq$Income2005, main="Scatterplot of Log Income Against Word",
    xlab="Word",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Word), col="red")
```



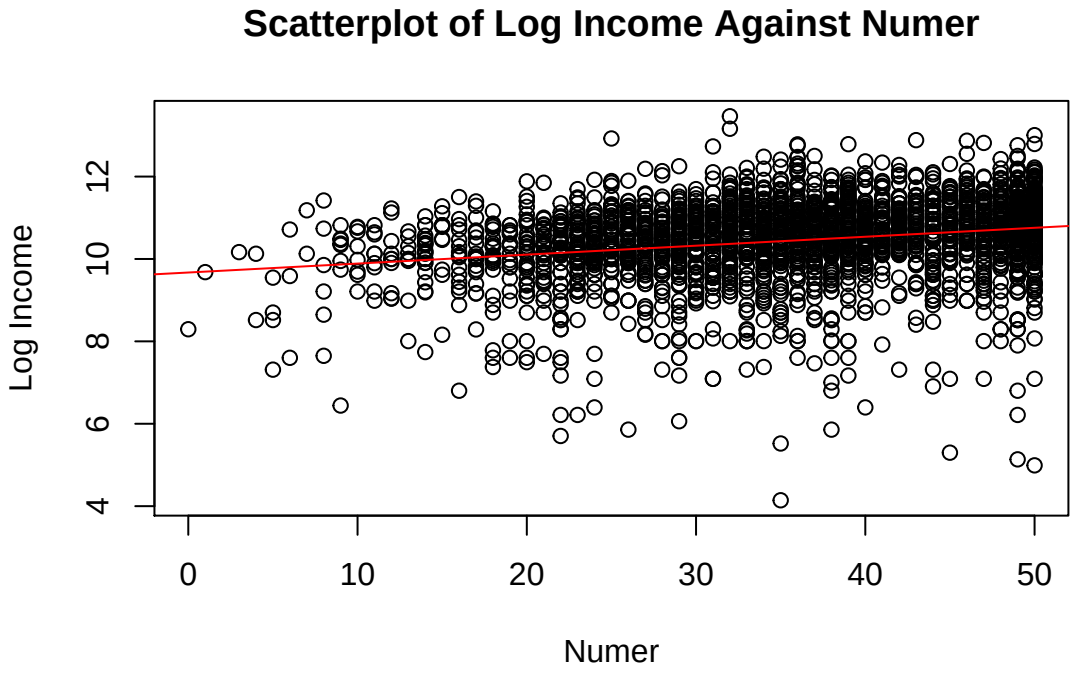
10. Parag and log income

```
plot(iq$Parag, iq$Income2005, main="Scatterplot of Log Income Against Parag",
    xlab="Parag",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Parag), col="red")
```



11. Numer and log income

```
plot(iq$Numer, iq$Income2005, main="Scatterplot of Log Income Against Numer",
    xlab="Numer",  ylab="Log Income")
abline(lm(iq$Income2005~iq$Numer), col="red")
```



As we can observe from the above plots, when other variables related to personal demographics and household environments are not adjusted, there are some mild positive associations between the the proxies of IQ and log income, although difference proxies of IQ might have different extents of effect on the log income. Moreover, we also observe that the variances are very large since the points on the scatterplots have very wide spread around the regression line.

ii. **Make a training data and testing data (approximately 2/3 observations as training data)**

We split the original data set `iq` to two data sets: `iq.train` is the training data, which consists of about 2/3 of the observations, and `iq.test` is the testing data, which consists of about 1/3 of the observations.

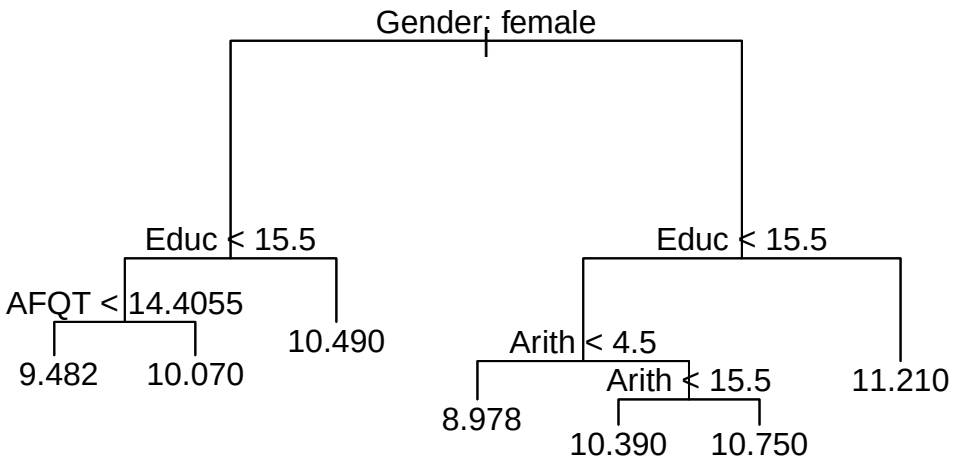
```
set.seed(233)
index <- sample(nrow(iq), (round(nrow(iq))/3))
iq.test <- iq[index, ]
iq.train <- iq[-index, ]
```

2. Single trees

i. **Build a reasonable single tree using all the variables to predict log income. Display the tree.**

Now, we build a reasonable single tree using all the variables to predict log income, and we also desplay the tree as below:

```
set.seed(2339)
fit1.single.full <- tree(Income2005~., iq.train)
plot(fit1.single.full)
text(fit1.single.full, pretty=0)
```



The related information of the tree that we have build are presented as below:

```
fit1.single.full$frame
```

```
##       var    n        dev      yval splits.cutleft splits.cutright
## 1  Gender 1723 1653.49524 10.446842             :a              :b
## 2    Educ  858  784.69270 10.141682          <15.5           >15.5
## 4    AFQT  590  439.38842  9.984782       <14.4055        >14.4055
## 8  <leaf>   83   83.68313  9.481717
## 9  <leaf>  507  331.26134 10.067138
## 5  <leaf>  268  298.80482 10.487095
## 3    Educ  865  709.65055 10.749533          <15.5           >15.5
## 6   Arith  612  421.54002 10.559972           <4.5            >4.5
## 12 <leaf>   11   37.07448  8.978444
## 13  Arith  601  356.44844 10.588918          <15.5           >15.5
## 26 <leaf>  266  167.02637 10.391697
## 27 <leaf>  335  170.86042 10.745517
## 7  <leaf>  253  212.92309 11.208076
```

ii. **Explain the predictions. Does that agree with our intuition?**

As we can see, the tree that we built patitions the predictors into different boxes using `Gender`, `AFQT`, `Arith`, and `Educ`. Sample mean in each box is then the predicted value for all the subjects in the box. Specifically, for a paticular subject, if the gender of the subject is female, and if she has received at least 15.5 years of education, then the predicted log income is 10.490; for a femaile with less than 15.5 years of education but scores at least 14.41 on the AFQT intelligence test, the predicted log income is 10.070; for a female with less than 15.5 years of education and scores less than 14.41 on the AFQT intelligence test, the predicted log income is 9.482. On the other hand, if the subject that we are interested in is male, and if he happens to receive at least 15.5 years of education, then the predicted log income is 11.210; for a male with less than 15.5 years of education and scores less than 4.5 on the arithmetic reasoning test, his predicted log income is 8.978; for a male with less than 15.5 years of education but scores at least 15.5 on the arithmetic reasoning test, his predicted income is 10.750; for a male with less than 15.5 years of education and whose score of the arithmetic reasoning test is between 4.5 and 15.5, the predicted log income is 10.390.

The tree that we conducted agrees with our intuition. First of all, men are generally paid more than women. Second, the number of years of eduction matters, in the sense that people receive more education are paid more. Finally, the IQ also matters, and people with higher scores on the proxy tests of IQ are expected to be paid more.

iii. **Can you detect gender bias against women from the tree built?**

Yes. gender bias against women can be detected from the tree built. Gender is the first predictor selected by the tree to split the data, and the predicted log income for males are higher than that for females. Therefore, gender plays a critical role in prediction using the tree that we built, and the gender bias against women does exist.

iv. **Do we have evidence that the high IQ's result in higher income? What is the testing errors for the tree you built?**

We also have evidence that the higher scores on some of the proxy tests of IQ will result in higher income, such as `arith` and `AFQT`, which have been displayed in the tree, and they also play important roles in spliting the data and make different predictions, especially among those people whose years of education are less than 15.5.

We would also like to calculate the testing errors for the tree that we built by using the testing data:

```
pred <- predict(fit1.single.full, iq.test[, -21])
mean((iq.test[,21]-pred)^2)
```

```
## [1] 0.7110214
```

3. Bagging with two bootstrap trees.

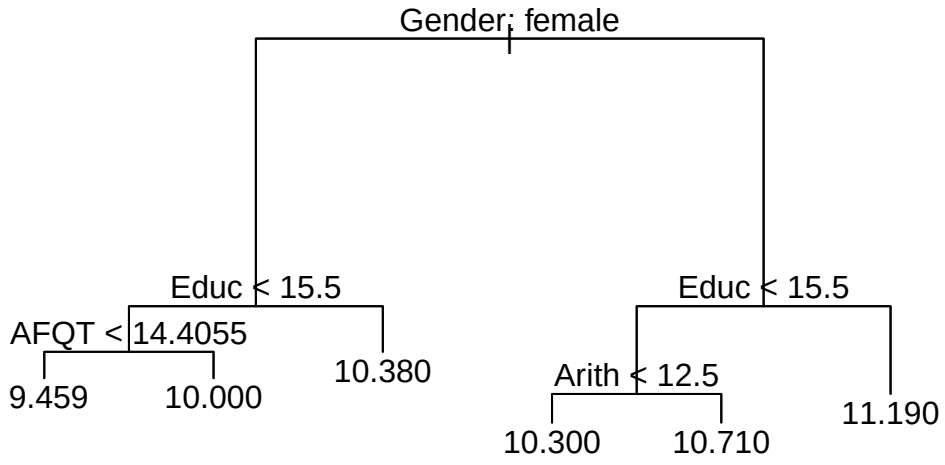
i. **Build and display two bootstrap trees. Commenting on the differences among the two trees.**

We start by building two bootstrap trees.

```
set.seed(233)
index1 <- sample(nrow(iq.train), nrow(iq.train), replace = T)
boot1 <- iq.train[index1, ]
set.seed(233)
tree1 <- tree(Income2005~., boot1)
set.seed(2333)
index2 <- sample(nrow(iq.train), nrow(iq.train), replace = T)
boot2 <- iq.train[index2, ]
set.seed(2333)
tree2 <- tree(Income2005~., boot2)
```

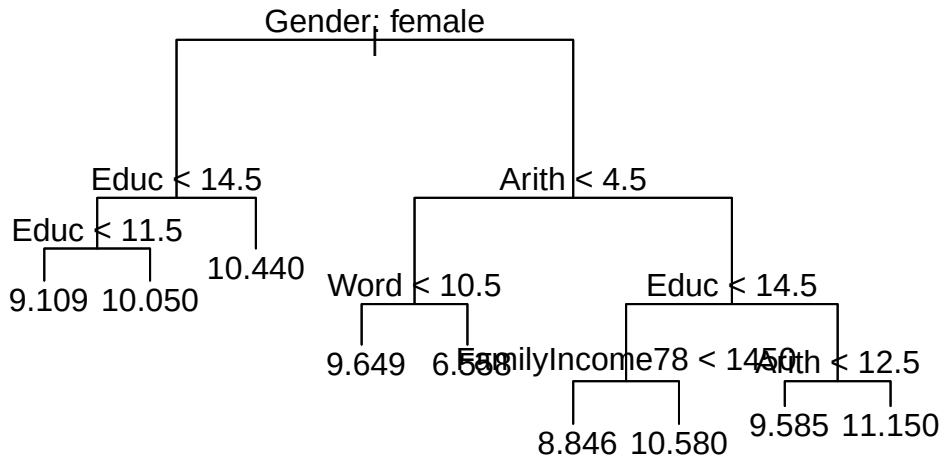
Then we plot the first tree as the following:

```
plot(tree1)
text(tree1, pretty=0)
```



The second tree is plotted as the following:

```
plot(tree2)
text(tree2, pretty=0)
```



In order to comment on the difference of the two trees, we can start by outputing the information about the first tree:

```
tree1$frame
```

```
##       var    n        dev      yval splits.cutleft splits.cutright
## 1  Gender 1723 1669.35476 10.417024             :a              :b
## 2    Educ  851  845.06713 10.062966          <15.5           >15.5
## 4    AFQT  601  457.50640  9.930476       <14.4055        >14.4055
## 8  <leaf>   81   51.70471  9.459276
## 9  <leaf>  520  385.01587 10.003875
## 5  <leaf>  250  351.64979 10.381471
## 3    Educ  872  613.49847 10.762556          <15.5           >15.5
## 6   Arith  611  340.51403 10.578572          <12.5           >12.5
## 12 <leaf>  193  137.17181 10.295831
## 13 <leaf>  418  180.78945 10.709120
## 7  <leaf>  261  203.88438 11.193262
```

As well as that for the second tree:

```
tree2$frame
```

```
##               var    n         dev      yval splits.cutleft splits.cutright
## 1          Gender 1723 1700.041379 10.432712             :a              :b
## 2            Educ  866  786.741113 10.169872          <14.5           >14.5
## 4            Educ  530  390.278904 10.001768          <11.5           >11.5
## 8          <leaf>   29   40.105654  9.108879
## 9          <leaf>  501  325.714650 10.053452
## 5          <leaf>  336  357.860120 10.435036
## 3           Arith  857  793.017487 10.698311           <4.5            >4.5
## 6            Word   13   42.784551  8.222317          <10.5           >10.5
```

```
## 12         <leaf>    7    8.521198  9.649033
## 13         <leaf>    6    3.391314  6.557815
## 7            Educ  844  669.308233 10.736449          <14.5          >14.5
## 14 FamilyIncome78  555  351.026545 10.546146          <1450          >1450
## 28         <leaf>   11   10.604182  8.846040
## 29         <leaf>  544  307.985513 10.580523
## 15          Arith  289  259.582870 11.101910          <12.5          >12.5
## 30         <leaf>    8   21.477836  9.584535
## 31         <leaf>  281  219.161224 11.145109
```

From the above output, we can notice that there are some differences between the two trees. First of all, the first tree has 6 terminal notes while the second tree has 9 terminal notes. Second, the first tree is formed by spliting `Gender`, `Educ`, `AFQT`, and `Arith`, while the second tree is formed by spliting `Gender`, `Educ`, `Arith`, `word`, and `FamilyIncome78`. Finally, in the first tree, the terminal nodes generally contain large numbers of observations (for example, the terminal node with the fewest observations still has 81 observations), but the terminal nodes of the second tree have relatively fewer observations (for example, there are 3 terminal nodes of the second tree with fewer than 10 observations).

ii. **Bag the two tree by taking average of the two bootstrap trees above. What is the testing error?**

Now we would like to bag the two tree by taking the average of the two bootstrap trees above, and the testing error is calculated as the following:

```
pred1 <- predict(tree1, iq.test[, -21])
pred2 <- predict(tree2, iq.test[, -21])
mean((iq.test[,21]-(pred1+pred2)/2)^2)
```

```
## [1] 0.7224834
```

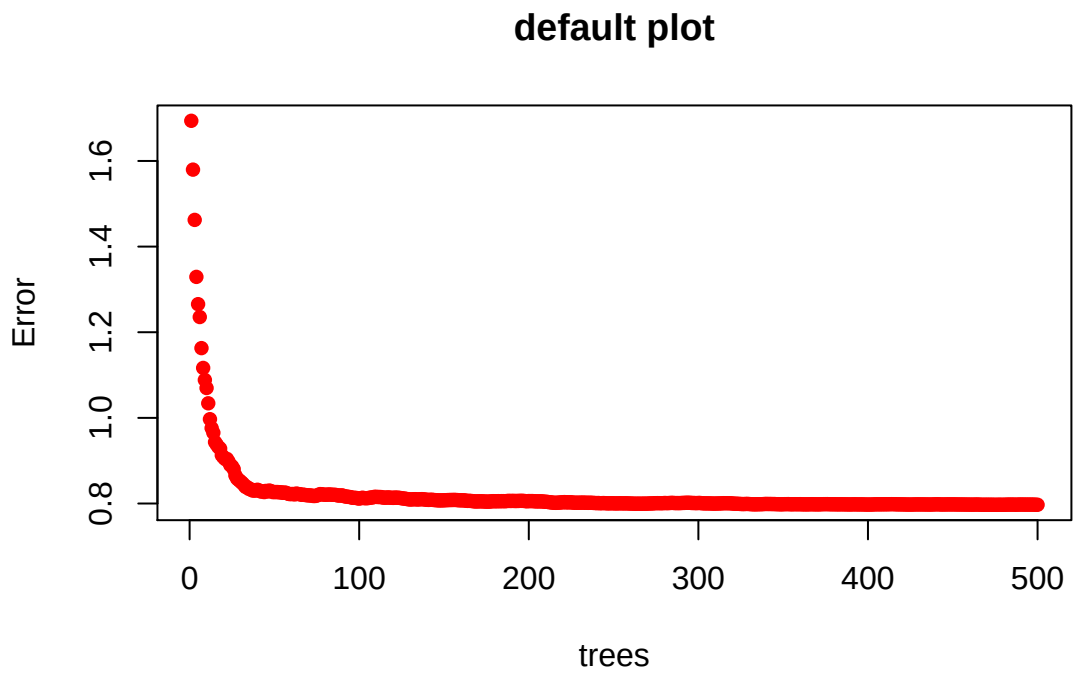
By comparing to the testing error calculated for a single tree in 2(iv), we conclude that they have very similar testing errors.

4. Build the best possible random forest

i. **Show the process how you tune the number of the trees and `mtry`**

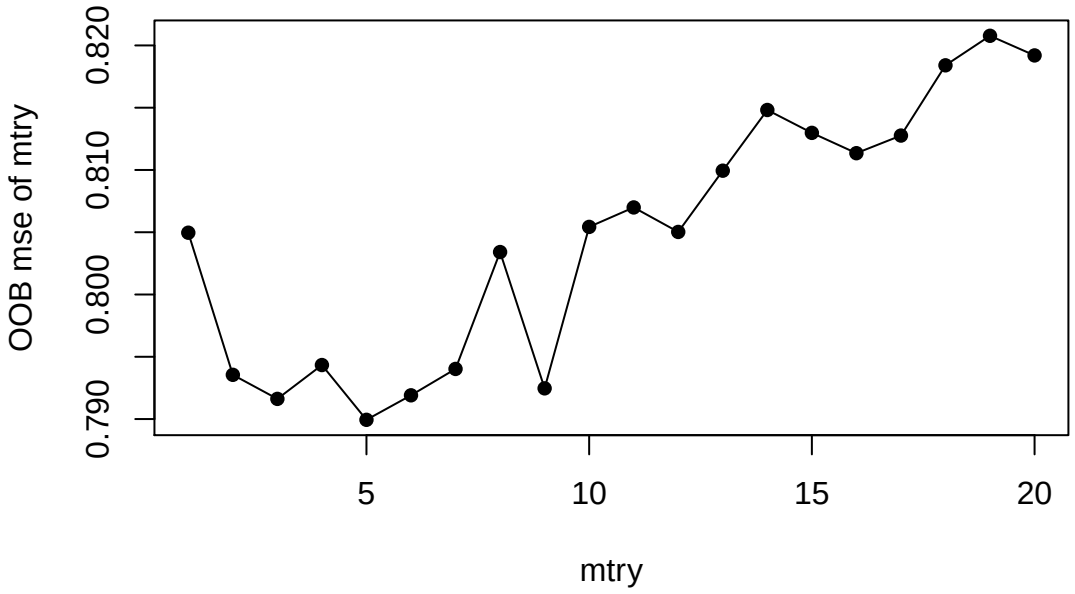
Since the total number of predictors is p=20, and the recommended `mtry` for regression tree is p/3=20/3=6 or 7. We can start by setting `mtry=7` and then tune the number of trees first.

```
fit.rf <- randomForest(Income2005~., iq.train, mtry=7, ntree=500)
plot(fit.rf, col="red", pch=16, type="p", main="default plot")
```



From the above plot, we can see that 350 trees would be sufficient to settle the OOB testing errors. Now, we fix `ntree=350`, and we would like to compare the OOB MSE to tune the `mtry`. Here we loop `mtry` from 1 to 20 (total number of available predictors) and return the tesing OOB errors.

```
rf.error.p <- 1:20
for (p in 1:20)
{
  set.seed(p*100)
  fit.rf <- randomForest(Income2005~., iq.train, mtry=p, ntree=350)
  rf.error.p[p] <- fit.rf$mse[350]
}
plot(1:20, rf.error.p, pch=16,
     xlab="mtry",
     ylab="OOB mse of mtry")
lines(1:20, rf.error.p)
```



From the above plot, the recommended choice of `mtry=7` look like a good choice, but choosing `mtry=5` looks even better.

Therefore, our final choice is `ntree=350` and `mtry=5`.

ii. **What is the testing error for your random forest?**

The testing error for the random forest is calculated as the following:

```
set.seed(700)
fit.rf.train <- randomForest(Income2005~., iq.train, mtry=5, ntree=350)
pred <- predict(fit.rf.train, iq.test[, -21])
mean((iq.test$Income2005-pred)^2)
```

```
## [1] 0.7062913
```

iii. **Comment on the pro and cons for the random forest comparing with the single tree you built earlier.**

There are some pros for the random forest. First of all, the random forest would potentially result in smaller variance compared to the single tree. Second, the random forest is formed by decorrelated trees. Third, random forest is easier to apply to the setting of high dimensional data.

There are also some cons with random forest. First of all, we need to tune `mtry`. If `mtry` is too small, then we would miss some key variables. On the other hand, if `mtry` is too large, then the correlation will not be reduced. Second, the random forest is harder to interpret compared to the single tree. Third, the random forest is more computationally expensive than the single tree.