Homework 5

Prakash Paudyal

Please do the following problems from the text book ISLR.

$\boldsymbol{1. Question 4.7.6\ pg\ 170}$

6. Suppose we collect data for a group of students in a statistics class with variables X1 =hours studied, X2 =undergrad GPA, and Y =receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0$ =???6, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1$..(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class. (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

In logistic regression, Estimated coefficients are given hence we substitute the values of coefficient in the logistic function to calculate the probability

x_1= Hours scheduled

x_2= Undergrad GPA

Y= receive on A

logistic regression, given estimated parameters

$\hat{\beta}_0$=-6

$\hat{\beta}_0$=-0.05

$\hat{\beta}_0$=-1, $ X_1$ = 40 hours x_2 =3.5 GPA

a) We can use this equation and Substitute the values of coefficients and predictors to calculate Probability.

$$\hat{p}\,(x) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}}$$

By substituting the given values on above equation, we get,

$$\hat{p}\,(x) = \frac{e^{-6+0.05*40+1*3.5}}{1+e^{-6+0.05*40+1*3.5}} = 0.3775$$

Hence probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class is 37.75%

b)

Here we are given probability $\hat{p}\ (x)$=0.5 and x_2=3.5 GPA, need to find the numbers of hours, Substituting this values in above logistic function.

$$\frac{1}{2} = \frac{e^{-6+0.05*x_1+1*3.5}}{1+e^{-6+0.05*x_1+1*3.5}}$$

or,

$$1+e^{-6+0.05*x_1+1*3.5} = 2e^{-6+0.05*x_1+1*3.5}$$

or,

$$1 = e^{-6+0.05*x_1+1*3.5}$$

Taking natural log(ln) on both sides

or,

$$0 = -6 + 0.05 * x_1 + 1 * 3.5$$

or,

$$x_1 = \frac{(6-3.5)}{0.05} = 50Hours$$

Hence, Student need to study 50hrs to have a 50% chance of getting A in the class.

$\mathbf{2.\ Question 4.7.7 pg 170}$

7. Suppose that we wish to predict whether a given stock will issue a dividend this year ("Yes" or "No") based on X, last year's percent profit. We examine a large number of companies and discover that the mean value of X for companies that issued a dividend was ¯X = 10, while the mean for those that didn't was ¯X = 0. In addition, the variance of X for these two sets of companies was = 36. Finally,80% of companies issued dividends. Assuming that X follows a normal distribution, predict the probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year. Hint: Recall that the density function for a normal random variable is

ANS: Given, $\hat{\sigma}^2$= 36

For k_{th} class for normal districution Bayes' theorem is

$$P_k(x) = \frac{\pi_k \cdot f_k(x)}{\sum_{l=1}^{k} \pi_l \cdot f_l(x)}$$

Because f(x) is normal then,

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k}} exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)$$

by substituting f(x) and given parameters in above equation we get.

$$P_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k}} exp(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2)}{\sum_{l=1}^{k} \pi_l \frac{1}{\sqrt{2\pi\sigma_l}} exp(-\frac{1}{2\sigma_l^2}(x-\mu_l)^2)}$$

or,

$$P_{yes}(x) = \frac{\pi_{yes} \frac{1}{\sqrt{2\pi\sigma_k}} exp(-\frac{1}{2\sigma_k^2}(x-\mu_{yes})^2)}{\pi_{yes} \frac{1}{\sqrt{2\pi\sigma_l}} exp(-\frac{1}{2\sigma_l^2}(x-\mu_{yes})^2) + \pi_{no} \frac{1}{\sqrt{2\pi\sigma_l}} exp(-\frac{1}{2\sigma_l^2}(x-\mu_{no})^2)}$$

μ_k *and* σ_k^2 **are the mean and variance parameters for the kth class which are given here, hence we can substitute these in above equation**

$$P_{yes} = \frac{0.80 * exp(-\frac{1}{2*36}(x-10)^2}{0.80 * exp(-\frac{1}{2*36}(x-10)^2 + 0.20 * exp(-\frac{1}{2*36}(x)^2}$$

or,

$$P_{yes}(X=4) = \frac{0.80 * exp(-\frac{1}{72}(4-10)^2}{0.80 * exp(-\frac{1}{72}(4-10)^2_+ 0.20 * exp(-\frac{1}{72}(4)^2}$$

so, $P_{yes}(X = 4)$=0.752

The probability that a company will issue a dividend this year given that its percentage profit was X = 4 last year is 75.2%

3. Continue from Homework #3 & 4 using the dataset from 4.7.10), fit a model for classification using the MclustDA function from the mclust-package.

```
## Package 'mclust' version 5.4
## Type 'citation("mclust")' for citing this R package in publications.

## ----------------------------------------------------
## Gaussian finite mixture model for classification
## ----------------------------------------------------
##
## MclustDA model summary:
##
##  log.likelihood   n df       BIC
##       -2129.439 985 10 -4327.804
##
## Classes   n Model G
##    Down 441     V 2
##    Up   544     V 2
##
## Training classification summary:
##
##       Predicted
## Class  Down  Up
##   Down   76 365
##   Up     70 474
##
## Training error = 0.4416244
##
## Test classification summary:
##
##       Predicted
## Class  Down Up
##   Down    5 38
##   Up      9 52
##
## Test error = 0.4134615
```

```
## Train Error IS:0.441624365482233
```

```
## Test Error IS:0.451923076923077
```

```
##       wTestClass
##        Down Up
##   Down    5  9
##   Up     38 52
```

```
## True Positive Rate is:0.577777777777778
```

```
## True Negative Rate is :0.357142857142857
```

Even though confusion matrix looks similar, classError() function is giving different test error than calculating by prediction. Hence, I calculated error by both methods.

```
Test error = 0.4134615
```

```
Test Error IS:0.451923076923077
```

```
i) What is the best model selected by BIC? What is the training error? What i
s the test error? Report the True Positive Rate and the True Negative Rate.
```

ModelName	trainError	testError	TPR	TNR
V	0.4416	0.4519	0.5778	0.3571
V	0.4416	0.4519	0.5778	0.3571

```
ii) Specify modelType="EDDA" and run MclustDA again. What is the best model s
elected by BIC? Find the training and test error rates. Report the True Posit
ive and True Negative Rate.
```

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## EDDA model summary:
##
##  log.likelihood   n df       BIC
##       -2204.237 985  3 -4429.152
##
## Classes   n Model G
##    Down 441     E 1
##    Up   544     E 1
##
## Training classification summary:
##
##       Predicted
## Class  Down  Up
##   Down   22 419
##   Up     20 524
##
## Training error = 0.4456853
##
## Test classification summary:
##
##       Predicted
## Class  Down Up
##   Down    9 34
##   Up      5 56
##
## Test error = 0.375
```

```
## Train Error IS:0.445685279187817
```

```
## Test Error IS:0.375
```

```
##       Predected
## Actual Down Up
##   Down    9 34
##   Up      5 56
```

```
## True Positive Rate is:0.918032786885246
```

```
##  True Negative Rate is :0.209302325581395
```

ModelName	trainError1	testError1	TPR1	TNR1
E	0.4457	0.375	0.918	0.2093
E	0.4457	0.375	0.918	0.2093

```
iii) Compare the results with Homework \#3 \& 4. Which method performed the b
est? Justify your answer.
```

I have chosen GLM and LDA from previous homework with lag2 as predictor variable with same data partition for all model. By comparing the Test error of GLM, LDA and MclustDA(EDDA) performance is similar and better than other models.

Models	Testerrors
GLM	0.3750
LDA	0.3750
MclustDA	0.4519
MclustDA(EDDA)	0.3750

4. Continue from Homework #3 & 4 using the dataset from 4.7.11). Fit a classification model using the MclustDA function from the mclust-package. Use the same training and test set from previous homework assignments.

Split the data into a training set and a test set.

fit MclustDa model for Auto data set

```
## ------------------------------------------------
## Gaussian finite mixture model for classification
## ------------------------------------------------
##
## MclustDA model summary:
##
##  log.likelihood   n df       BIC
##       -5535.636 294 88 -11571.43
##
## Classes   n Model G
##       0 149   VVE 5
##       1 145   VEE 5
## 
## Training classification summary:
## 
##      Predicted
## Class   0   1
##     0 142   7
##     1   4 141
## 
## Training error = 0.03741497 
## 
## Test classification summary:
## 
##      Predicted
## Class  0  1
##     0 42  5
##     1  5 46
## 
## Test error = 0.1020408
```

```
## Train Error IS:0.0374149659863946
```

```
## Test Error IS:0.102040816326531
```

```
##       Predected
## Actual  0  1
##      0 42  5
##      1  5 46
```

```
## True Positive Rate is:0.901960784313726
```

```
## True Negative Rate is :0.893617021276596
```

i) What is the best model selected by BIC? What is the training error? What is the test error? Report the True Positive Rate and the True Negative Rate.

ModelName	trainError	testError	TPR	TNR
VVE	0.0374	0.102	0.902	0.8936
VEE	0.0374	0.102	0.902	0.8936

ii) Specify modelType="EDDA" and run MclustDA again. What is the best model selected by BIC? Find the training and test error rates. Report the True Positive and True Negative Rate.

```
## ------------------------------------------------ 
## Gaussian finite mixture model for classification
## ------------------------------------------------
## 
## EDDA model summary:
## 
##  log.likelihood   n df       BIC
##       -5784.354 294 22 -11693.75
## 
## Classes   n Model G
##       0 149   VVE 1
##       1 145   VVE 1
## 
## Training classification summary:
## 
##      Predicted
## Class   0   1
##     0 132  17
##     1  10 135
## 
## Training error = 0.09183673 
## 
## Test classification summary:
## 
##      Predicted
## Class  0  1
##     0 40  7
##     1  3 48
## 
## Test error = 0.1020408

## Train Error IS:0.0918367346938776

## Test Error IS:0.102040816326531

##       Predected
## Actual  0  1
##      0 40  7
##      1  3 48

## True Positive Rate is:0.941176470588235

## True Negative Rate is :0.851063829787234
```

ModelName	trainError	testError	TPR	TNR
VVE	0.0918	0.102	0.9412	0.8511
VVE	0.0918	0.102	0.9412	0.8511

```
iii) Compare the results with Homework \#3 \& 4. Which method performed the b
est? Justify your answer.
```

I choose QDA model from previous homework with lowest test error. I used same data partition method and predictor variable as before (mpg01 ~ displacement+horsepower+weight+year). All THE methods, QDA, MclustDA and MclustDA(EDDA) produce same test error.

Methods	Test_error
QDA	0.102
MclustDA	0.102
MclustDA(EDDA)	0.102

5. Read the paper “Who Wrote Ronald Reagan’s Radio Addresses?” posted on D2L. Write a one page (no more, no less) summary.

Title: Who wrote Ronald Reagan’s radio addresses?

Authors: Edoardo M. Airoldi, Annelise G. Anderson, Stephen E. Fienberg, and Kiron K. Skinner

This paper aims to determine the author of 312 out of around 1000 radio broadcasts that were by Ronald Reagan during his presidency campaign from 1975 to 1979, for which no direct evidence of authorship is available. Since Peter Hannaford, who used to assist Reagan in drafting texts for radio addresses was in service from 1976-1979, speculations were being made about the person who could have helped Reagan in this task. In order to get a better understanding of the situation, the authors used statistical techniques to learn a way to distinguish between the writing styles of Reagan and Hannaford. They also focused on the stylistic differences Reagan and his other collaborators in order to predict the authors of the speeches that he delivered at different points in time.

One of the most crucial issues that the authors had to deal with was the striking difference in the number of known speeches that were available for each author and the phase of word selection. The original dataset comprised of 679 speeches that had been drafted by Regan in his own handwriting but there were only 39 speeches that had been drafted by his close collaborators. However, with the help of Kiron Skinner and Annelise Anderson, the authors looked into the Regan files and found 30 newspaper columns that were published under Reagan’s name but had been originally drafted by Peter Hannaford. After coding these 30 articles, finally a set of 69 texts that were drafted by Reagan’s collaborators was prepared. The authors explored a wide range of classification models as well as the Negative Binomial model and fully Bayesian models for word counts. As a starting point for their modeling choices of word count data, they used the study of “The Federalist” papers authored by Frederick Mosteller and David Wallace. This study not only helped them with figuring out the heuristics for selecting features, but also provided ideas for dealing with issues that arise for Negative-Binomial counts when the texts that are being sampled have different lengths.

The authors produced separate sets of predictions using the most accurate classification methods and the full Bayesian models for the 314 speeches whose authors were not known. All of the predictions agreed on 135 of the unknown speeches while the fully Bayesian models agreed on 289 of them. They also provided separate models for the speeches in 1975 and those in 1976-1979 and obtained stable predictions on speeches that were given about various topics in different years. An assumption that was made in all the models was that the words are independent of each other. Even though this assumption does not always hold in general, since the models rely on our-of-sample cross-validation, the results that were obtained are not overstatements or misrepresentations. The cross-validated accuracy of the fully Bayesian model based on the Negative-Binomial distribution for word counts was above 90% in all the cases.

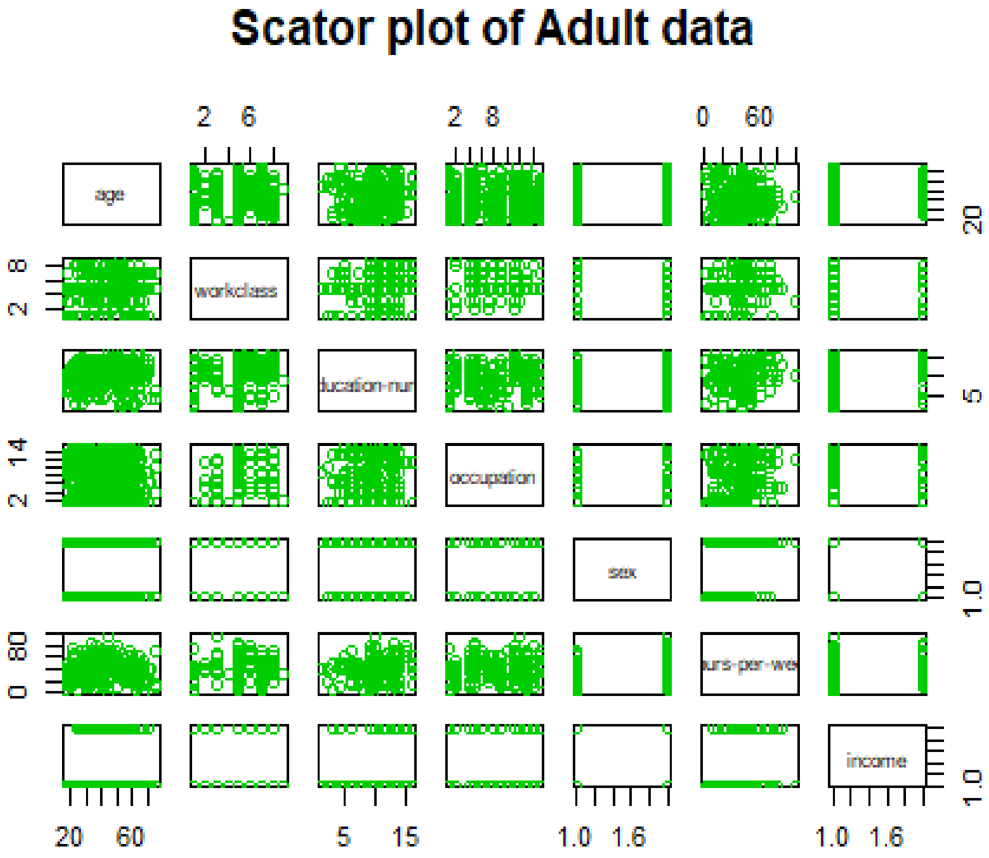
Based on these analyses, the authors were able to conclude that in the year 1975, Regan drafted 77 of his speeches, while his collaborators drafted 71 of them. Similarly, over the years 1976-1979, Regan drafted 9 speeches while Hannaford drafted 74.

6. Last homework you chose a dataset from this website. Please do some initial exploration of the dataset. If you don't like the dataset you chose you may change it with another. It has to be a new dataset that we haven't used in class. Please report the analysis you did and discuss the challenges with analyzing the data.

ANS:
The dataset that I have chosen is the "Census Income" dataset, also known as "Adult" dataset. It consists of 14 attributes and 48842 data samples. Based on these attributes, the task is to predict whether the income exceeds $50K/year.
First step I did try to explore the numbers of variables and their type by using str().Then I check if I need all variables to analysis the data or not. Some variables were not useful like capital-loss, capital-profit because few observations only have their value.
There were more variables and some of them with high levels of factor, it is challenging to identify the usefulness of variables. I removed some variable by assuming that they are not useful predictor e.g. marital status, relationship.
Then I plot the scatter plot of data of the data and tried to find some relationship between variables. There was no significant correlation between the predictor variables. From the scatter plot, Age, sex , education number are good predictor for income.

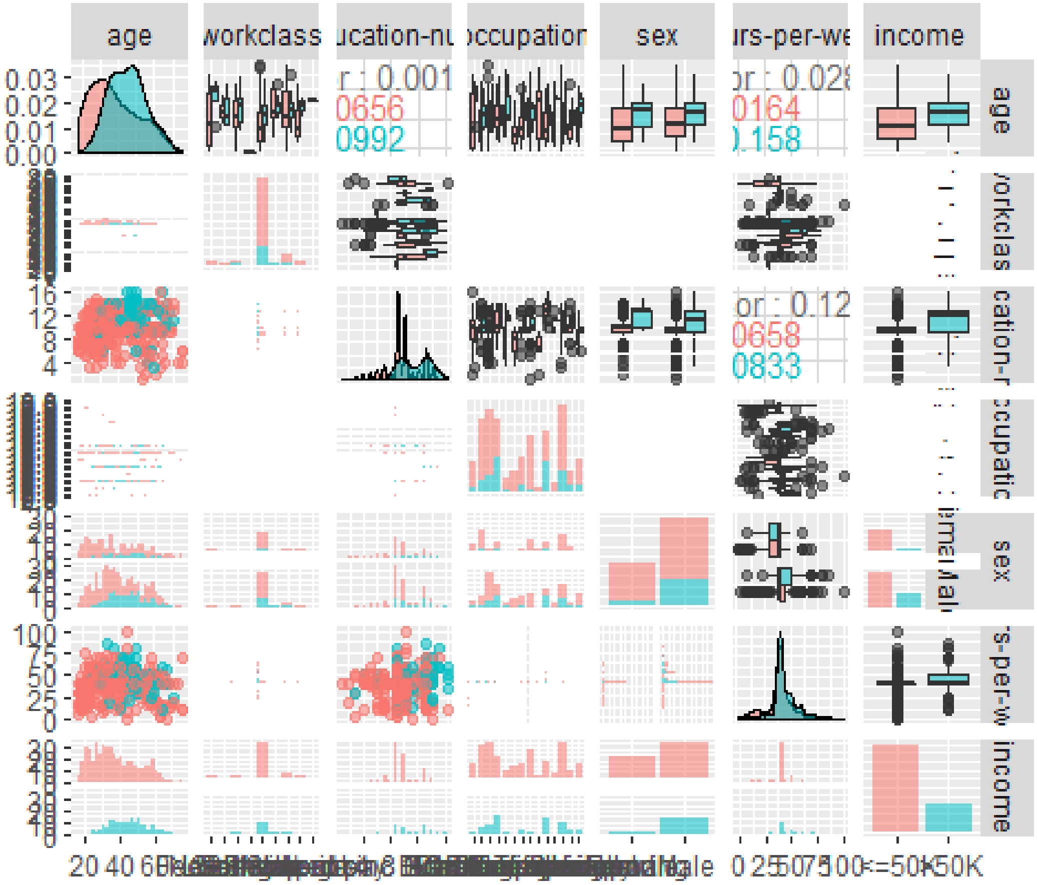
```
##       age  workclass fnlwgt  education education-num      marital-status
## 2259   29    Private 241431    Masters            14  Married-civ-spouse
## 26627  25    Private 234665  Assoc-voc            11  Married-civ-spouse
## 30691  39    Private  91367    Masters            14  Married-civ-spouse
## 8771   27  Local-gov 124680    Masters            14       Never-married
## 5514   55    Private 231738    HS-grad             9            Divorced
## 1104   41    Private 118619  Bachelors            13  Married-civ-spouse
##              occupation   relationship   race     sex capital-gain
## 2259              Sales        Husband  White    Male         7298
## 26627      Craft-repair        Husband  White    Male            0
## 30691   Exec-managerial        Husband  White    Male            0
## 8771     Prof-specialty  Not-in-family  White  Female            0
## 5514              Sales  Not-in-family  White  Female            0
## 1104    Exec-managerial        Husband  Black    Male            0
##       capital-loss hours-per-week native-country income
## 2259             0             40  United-States   >50K
## 26627            0             45  United-States  <=50K
## 30691         1848             45  United-States   >50K
## 8771             0             40  United-States  <=50K
## 5514             0             40        England  <=50K
## 1104             0             50  United-States  <=50K

## 'data.frame':    651 obs. of  15 variables:
##  $ age           : int  29 25 39 27 55 41 37 37 17 61 ...
##  $ workclass     : Factor w/ 9 levels " ?"," Federal-gov",..: 5 5 5 3 5 5
7 3 5 5 ...
##  $ fnlwgt        : int  241431 234665 91367 124680 231738 118619 282461 39
7877 316929 69867 ...
```

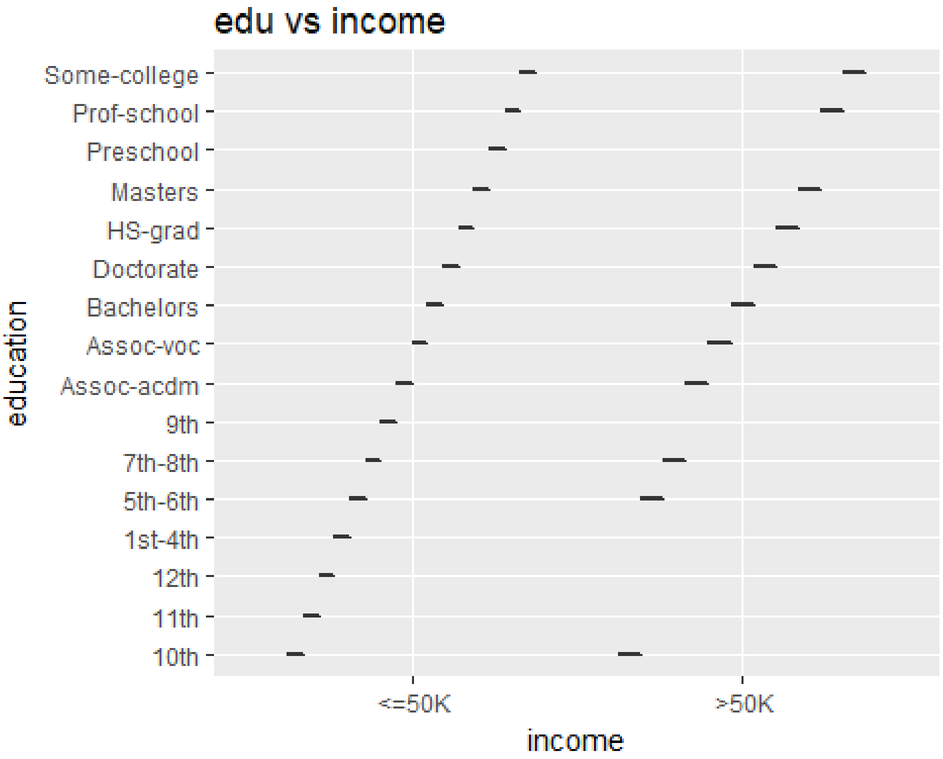
```
##  $ education     : Factor w/ 16 levels " 10th"," 11th",..: 13 9 13 13 12 1
0 16 10 3 12 ...
##  $ education-num : int  14 11 14 14 9 13 10 13 8 9 ...
##  $ marital-status: Factor w/ 7 levels " Divorced"," Married-AF-spouse",..:
3 3 3 5 1 3 3 3 5 3 ...
##  $ occupation    : Factor w/ 15 levels " ?"," Adm-clerical",..: 13 4 5 11
13 5 5 2 7 5 ...
##  $ relationship  : Factor w/ 6 levels " Husband"," Not-in-family",..: 1 1
1 2 2 1 1 1 4 1 ...
##  $ race          : Factor w/ 5 levels " Amer-Indian-Eskimo",..: 5 5 5 5 5
3 5 5 5 5 ...
##  $ sex           : Factor w/ 2 levels " Female"," Male": 2 2 2 1 1 2 2 2 2
2 ...
##  $ capital-gain  : int  7298 0 0 0 0 0 0 0 0 0 ...
##  $ capital-loss  : int  0 0 1848 0 0 0 0 0 0 0 ...
##  $ hours-per-week: int  40 45 45 40 40 50 60 40 20 40 ...
##  $ native-country: Factor w/ 42 levels " ?"," Cambodia",..: 40 40 40 40 10
40 40 40 40 40 ...
##  $ income        : Factor w/ 2 levels " <=50K"," >50K": 2 1 2 1 1 1 1 2 1
2 ...
```

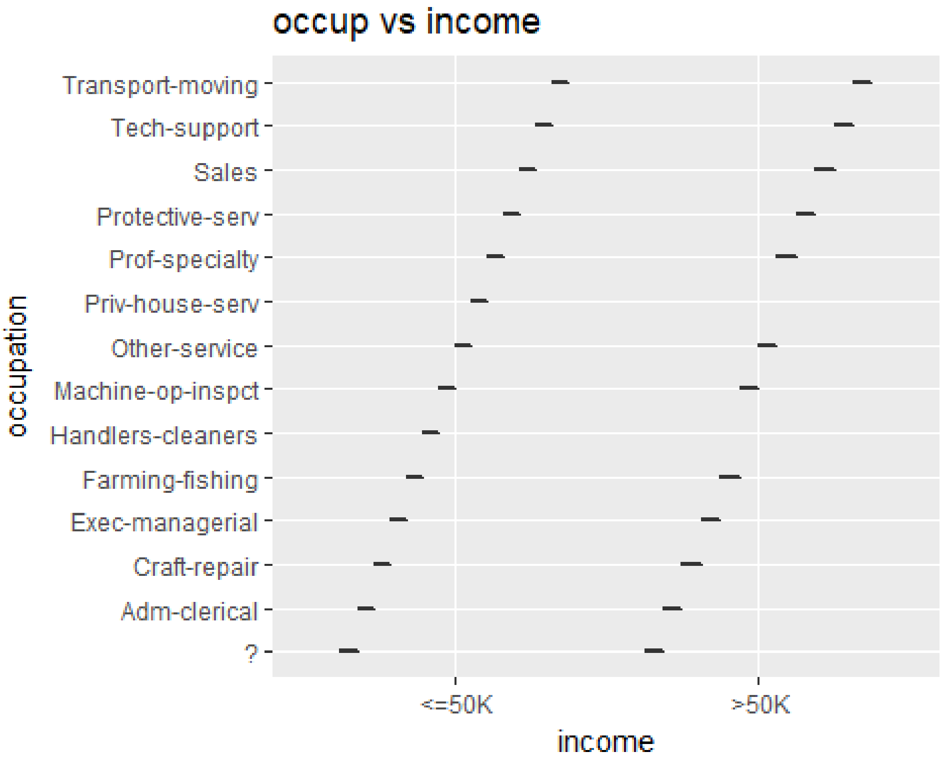

Scator plot of Adult data

```
##
##  <=50K   >50K
##    496    155
```

ggplot




edu vs income
education
Some-college
Prof-school
Preschool
Masters
HS-grad
Doctorate
Bachelors
Assoc-voc
Assoc-acdm
9th
7th-8th
5th-6th
1st-4th
12th
11th
10th
<=50K
>50K
income


occup vs income
occupation
Transport-moving
Tech-support
Sales
Protective-serv
Prof-specialty
Priv-house-serv
Other-service
Machine-op-inspct
Handlers-cleaners
Farming-fishing
Exec-managerial
Craft-repair
Adm-clerical
?
<=50K
>50K
income