

DAT 500S Module 3

Textbook Quiz:

Exercise 4.7 Q1 & Q4

Group Assignment:

Exercise 4.7 Q11(-part e) & Q12

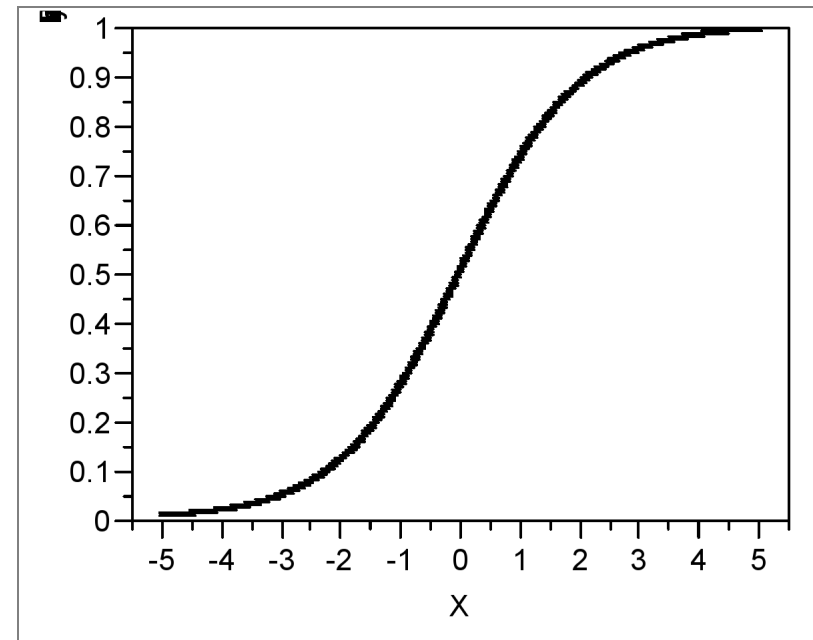
Individual Assignment:

Exercise 4.7 Q10

Module 3 - Logistic Regression

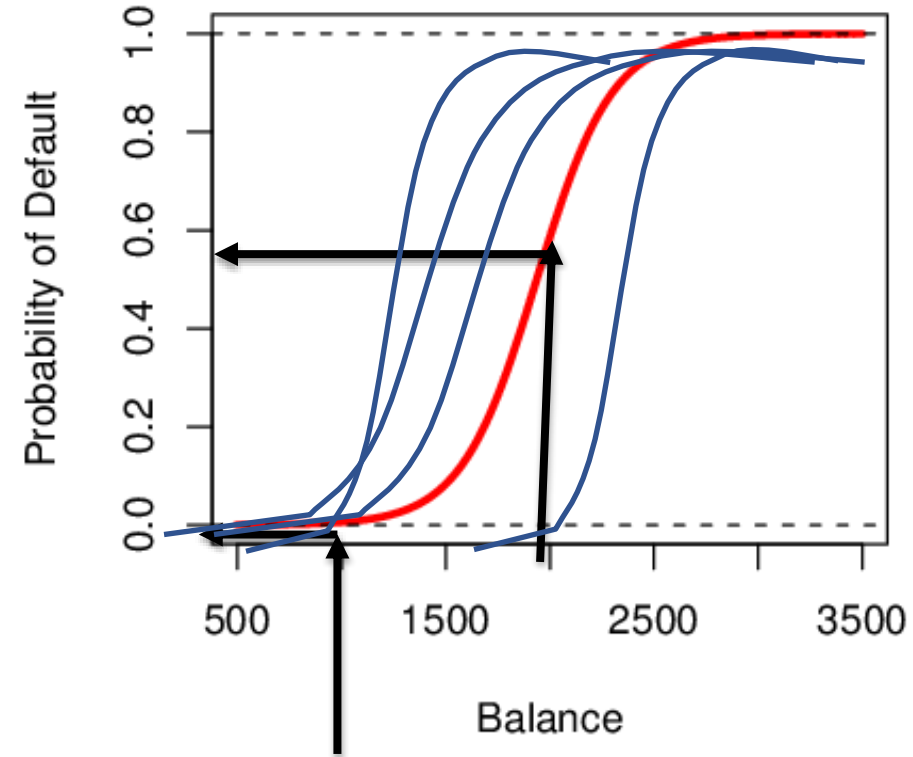
- Instead of trying to predict Y , let's try to predict $P(Y = 1)$, i.e., the probability a customer buys Citrus Hill (CH) juice.
- Thus, we can model $P(Y = 1)$ using a function that gives outputs between 0 and 1.
- We can use the **logistic function**
- Logistic Regression!

$$p(X) = P(Y=1 / X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$



Module 3 - Logistic Regression

- The probability of default is close to zero for a balance of \$1000.
- The probability of default is close to 0.6 for a balance of \$2000.



Bayes' Theorem


.



No
Image

LDA from Bayes' Theorem

- The most common model for $f_k(x)$ is the Normal Density

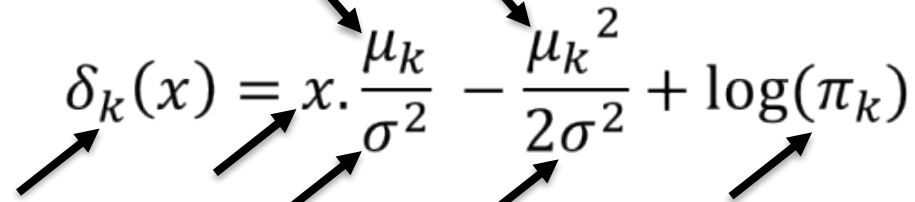
$$f_k(x) = \frac{1}{\sigma_k \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma_k^2} (x - \mu_k)^2\right)$$


- Assume, $\sigma_1 = \sigma_2 = \dots = \sigma_K = \sigma$

$$p_k(x) = \frac{\pi_k \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} (x - \mu_k)^2\right)}{\sum_{l=1}^K \pi_l \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2} (x - \mu_l)^2\right)}$$

Discriminating Function

- Assign the observation to the class for which $\delta_k(x)$ is largest

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$
The diagram shows the equation $\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$ with arrows pointing from its components to the text in the following bullet point. Arrows point from $\delta_k(x)$ to the first $\delta_k(x)$, from x to the x , from $\frac{\mu_k}{\sigma^2}$ to $\mu_1 = \mu_2$, from $\frac{\mu_k^2}{2\sigma^2}$ to $\mu_1^2 = \mu_2^2$, and from $\log(\pi_k)$ to $\pi_1 = \pi_2$.

- For example, if $K = 2$ and $\pi_1 = \pi_2$, then assign an observation to class 1 if $\delta_1(x) > \delta_2(x)$

-
- i.e. if $2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$

-
- Draw the boundary between classes at $\delta_1(x) = \delta_2(x)$

Use Training Data set for Estimation

$$\longrightarrow \hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$$

$$\longrightarrow \hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^K \sum_{i: y_i = k} (x_i - \hat{\mu}_k)^2$$

$$\longrightarrow \hat{\pi}_k = n_k / n.$$

Quiz: Exercise 4.7 Q1

Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent

$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \quad (4.2)$$

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X} \quad (4.3)$$

$$\begin{aligned} 1 - p(X) &= 1 - \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \\ &= \frac{1 + e^{\beta_0 + \beta_1 X} - e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{\beta_0 + \beta_1 X}} \\ \frac{p(X)}{1 - p(X)} &= \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} \times (1 + e^{\beta_0 + \beta_1 X}) = e^{\beta_0 + \beta_1 X} \end{aligned}$$

Quiz: Exercise 4.7 Q4 – Curse of Dimensionality

(a)

One-feature observations X uniformly distributed on $[0, 1]$.

Associated with each observation is a response value.

Make prediction with 10% nearest neighbors. \rightarrow [test $x+0.05$, test $x-0.05$]

On average, what fraction of observations will be used in prediction?

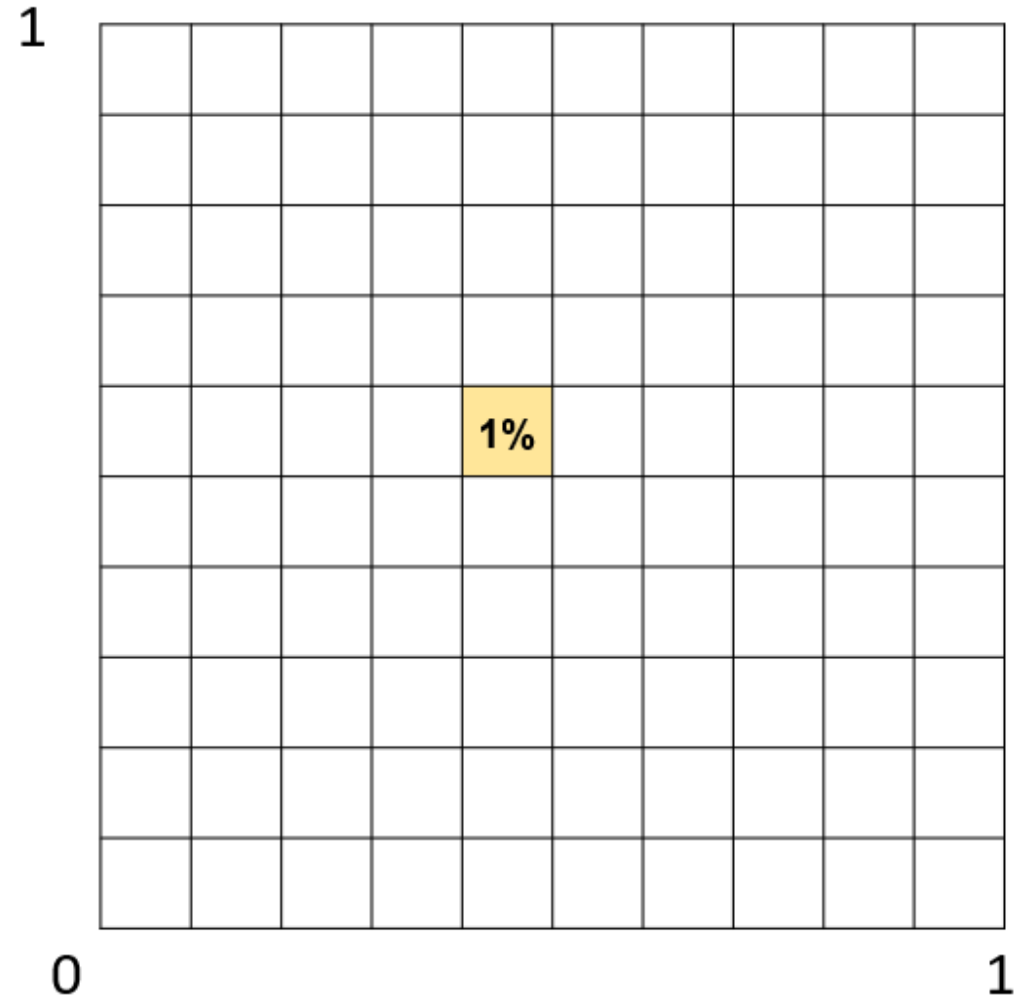
- On average, 10% of total observations will be used in prediction



Quiz: Exercise 4.7 Q4 – Curse of Dimensionality

(b)

- Two-feature observations (X_1, X_2) uniformly distributed on $[0,1] \times [0,1]$.
- Make prediction with 10% range on each dimension.
- $[x_1 + 0.05, x_1 - 0.05] \times [x_2 + 0.05, x_2 - 0.05]$
- $(0.1)^2 = 0.01 = 1 \%$



Quiz: Exercise 4.7 Q4 – Curse of Dimensionality

(c)

100-feature observations X_1, \dots, X_{100} all uniformly distributed on $[0, 1]$.

Make prediction with 10% range on each dimension.

-> $[x_1+0.05, x_1-0.05] \times [x_2+0.05, x_2-0.05] \dots [x_{100}+0.05, x_{100}-0.05]$

On average, what fraction of observations will be used in prediction?

- On each dimension: use 10% of range
- 100 dimensions all together: $(0.1)^{100}$

Quiz: Exercise 4.7 Q4 – Curse of Dimensionality

(d)

Using your answers to parts (a)–(c), argue that a drawback of kNN when p is large is that there are very few training observations “near” any given test observation.

- As p increases, one has to go far away to find neighbors for the candidate point for which prediction is made. Those far away neighbors won't necessarily reflect the same characteristics of the candidate point.

Quiz: Exercise 4.7 Q4 – Curse of Dimensionality

(d)

p-dimensional hypercube:

centered around the test observation

contains 10% of the training observations on average.

For $p = 1, 2$, and 100, what is the length of each side of the hypercube?

- Each dimension has range $[0,1]$
- $\text{Length}^{\text{Dimension}} = \text{Area}$
 $\Rightarrow \text{Length} = \text{Area}^{\frac{1}{\text{Dimension}}}$
- $p=1$: length = $0.1^{\frac{1}{1}} = 0.10$
- $p=2$: length = $0.1^{\frac{1}{2}} \approx 0.31$
- $p=100$: length = $0.1^{\frac{1}{100}} \approx 0.98$
- When dimensionality is high to ensure a large enough training sample size we need to increase the range of sampling on each dimension (side length)

Group Assignment: Problem 11

```
Auto = read.csv("Auto.csv",header = T, na.strings = "?")  
fix(Auto)  
Auto = na.omit(Auto)
```

- (a) Binary variable mpg01: contains 1 if mpg contains a value above its median, and 0 if mpg contains a value below its median

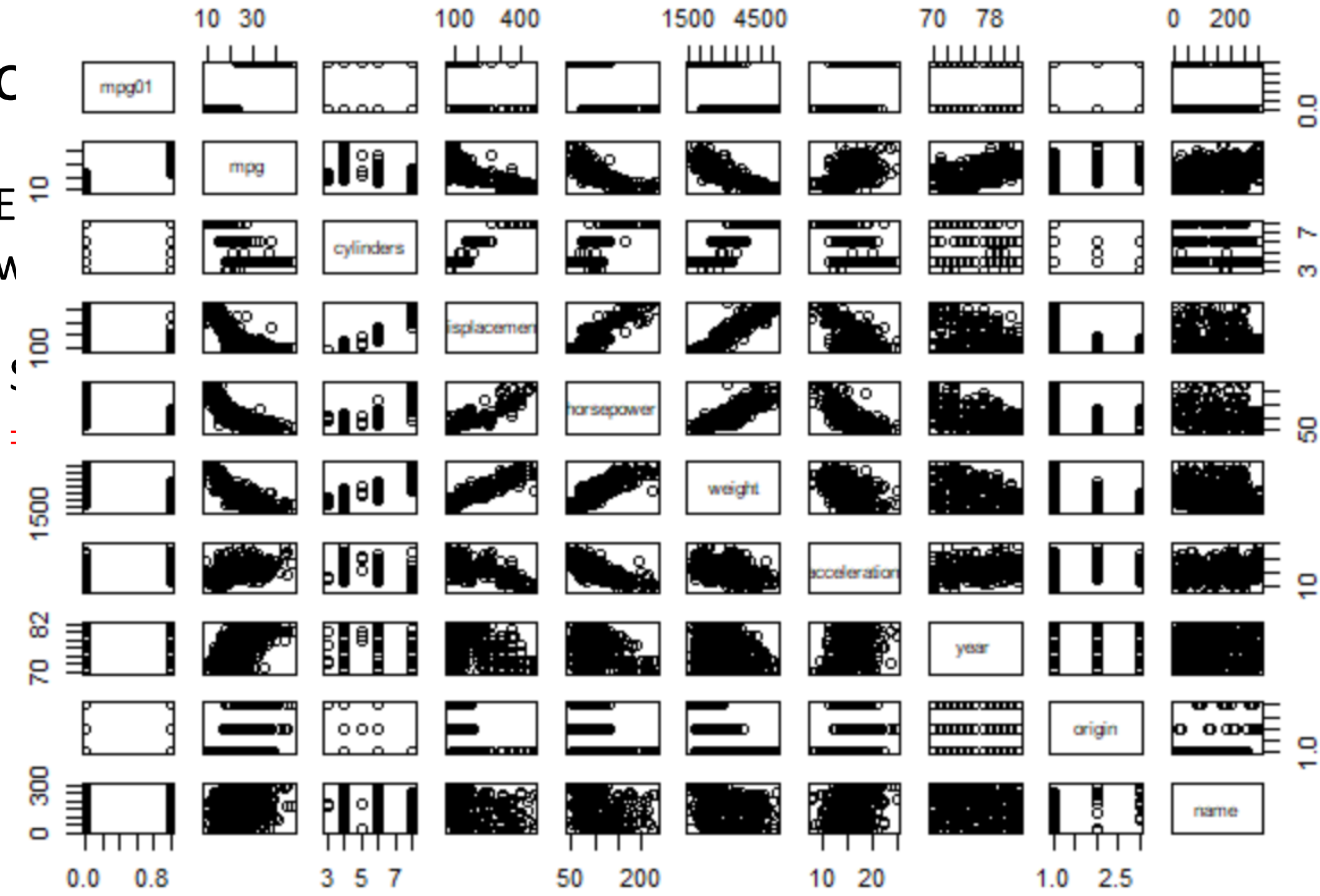
```
⇒mpg01<-ifelse(Auto$mpg>median(Auto$mpg),1,0)
```

Create new dataframe:

```
⇒Auto<-data.frame(cbind(mpg01,Auto))
```

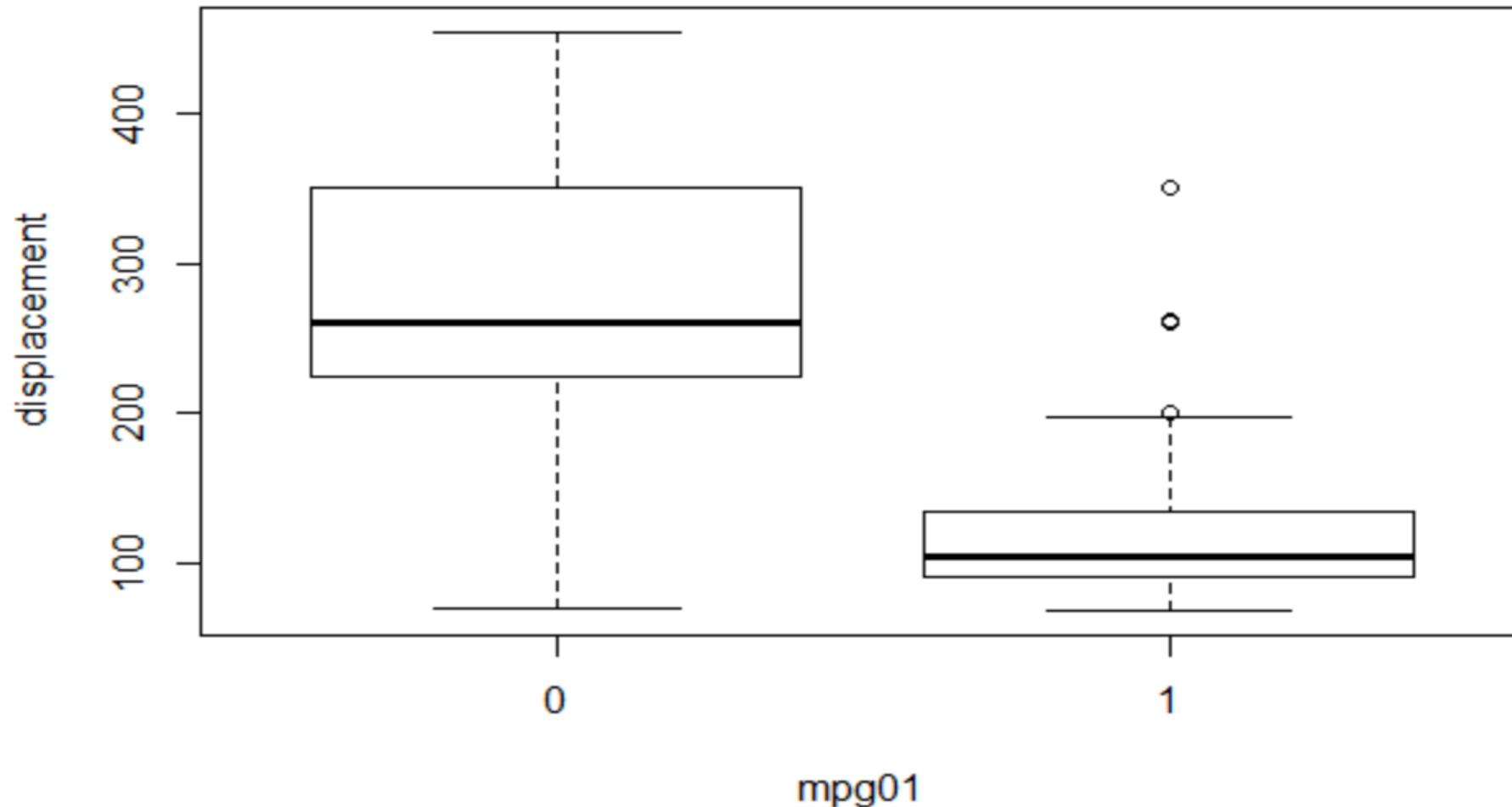
Grc

(b) E
betw



Group Assignment: Problem 11

(b) Explore the data graphically in order to investigate the association between mpg01 and the other features.



Group Assignment: Problem 11

(c) Split the data into a training set and a test set

Set random seed

⇒ `set.seed(100)`

Split train & test

⇒ `train = sample(nrow(Auto), nrow(Auto)/2, replace = FALSE)`

...

Group Assignment: Problem 11

(d) Perform LDA on the training data in order to predict mpg01

Use displacement, weight, year

⇒

```
library(MASS)
```

```
lda.fit<-lda(mpg01~weight+displacement+year,data=Auto[train,])
```

```
lda.pred<-predict(lda.fit,Auto[-train,])
```

Test error

⇒

```
table.lda<-table(lda.pred$class,Auto[-train,]$mpg01)
```

```
mean(lda.pred$class!=Auto[-train,]$mpg01)
```

Group Assignment: Problem 11

(f) Perform logistic regression on the training data in order to predict mpg01

Use displacement, weight, year

⇒

```
glm.fit<-
```

```
glm(mpg01~weight+displacement+year,data=Auto[train,],family=binomial)
```

```
glm.probs<-predict(glm.fit, Auto[-train,],type="response")
```

set threshold at 0.5, and get test error

```
⇒glm.pred = ifelse(glm.probs>0.5,1,0)
```

```
table.glm<-table(glm.pred, Auto[-train,]$mpg01)
```

```
mean(glm.pred!=Auto[-train,]$mpg01)
```

Group Assignment: Problem 11

(g) Perform kNN on the training data, with several values of K, in order to predict mpg01.

Use displacement, weight, year (column 4,6,8)

⇒ `library(class)`

`knn.pred<-knn(Auto[train,c(4,6,8)],Auto[-train,c(4,6,8)],
Auto[train,]$mpg01, k=1)`

test error ⇒ `mean(knn.pred!=Auto[-train,]$mpg01)`

Try different k (for loop)

Group Assignment: Problem 12 – Function

(a) Power(): compute & print 2^3

Define function:

```
⇒ Power = function(){  
    print(2^3)  
}
```

Call function

```
⇒ Power()
```

Group Assignment: Problem 12 – Function

(b) Power2(x,a): compute & print x^a

Define function:

```
⇒ Power2 = function(x,a){  
    print(x^a)  
}
```

(c) Call Power2(x,a) to compute: 10^3 , 8^{17} , 131^3

10^3 :

⇒ Power2(10,3)

8^{17} :

⇒ Power2(8,17)

131^3 :

⇒ Power2(131,3)

Group Assignment: Problem 12 – Function

(d) Power3(): with return value

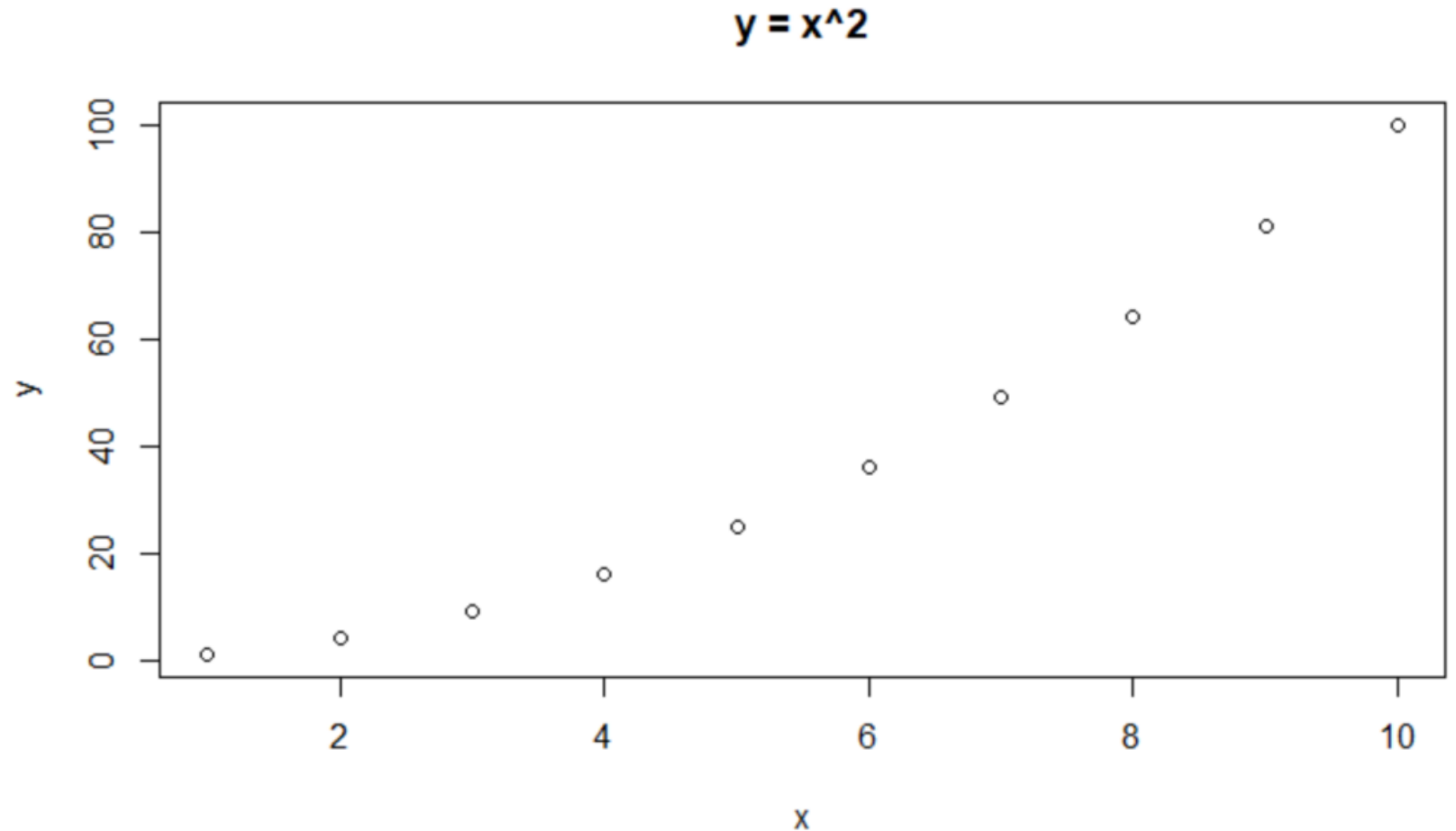
⇒

```
Power3 = function(x,a){  
    result = x^a  
    return(result)  
}
```

Group Assignment: Problem 12 – Function

(e) Plotting $y = x^2$

```
plot(  
  1:10,  
  Power3(1:10,2),  
  main = "y = x^2",  
  xlab = "x",  
  ylab = "y"  
)
```

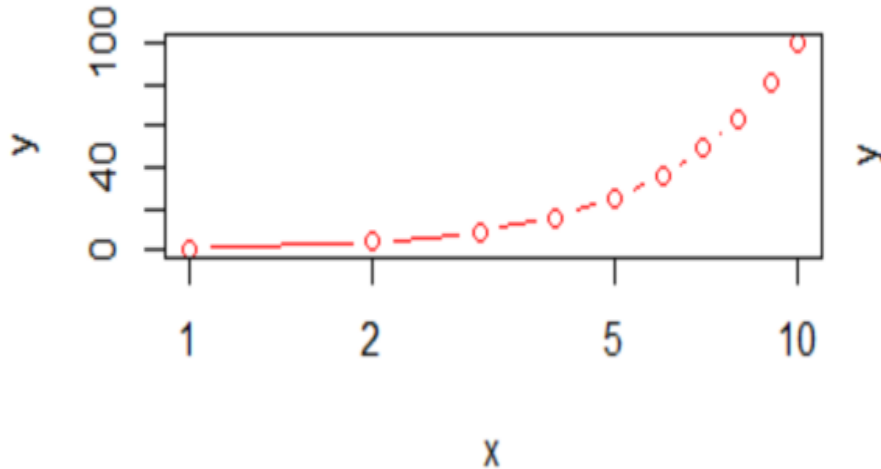


Group Assignment: Problem 12 – Function

(e) Plotting $y = x^2$, with log-scaling

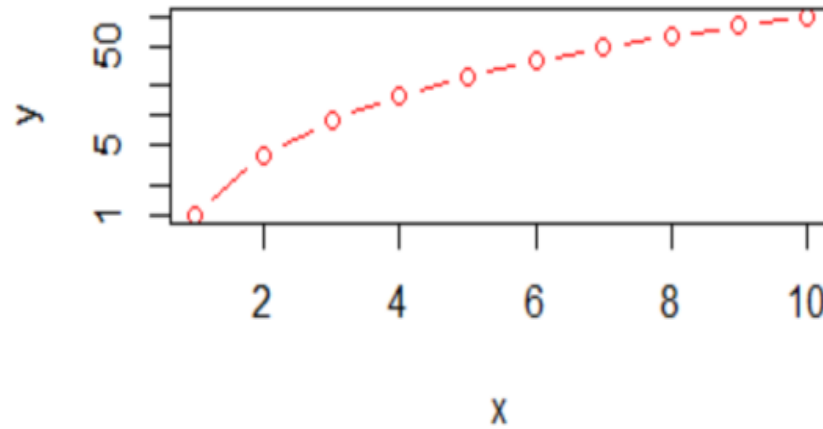
```
plot(  
  1:10,  
  Power3(1:10,2),  
  log = "x",  
  main = 'y = x^2, log="x"',  
  xlab = "x",  
  ylab = "y",  
  type = "b",  
  col = "red"  
)
```

$y = x^2, \text{log}="x"$



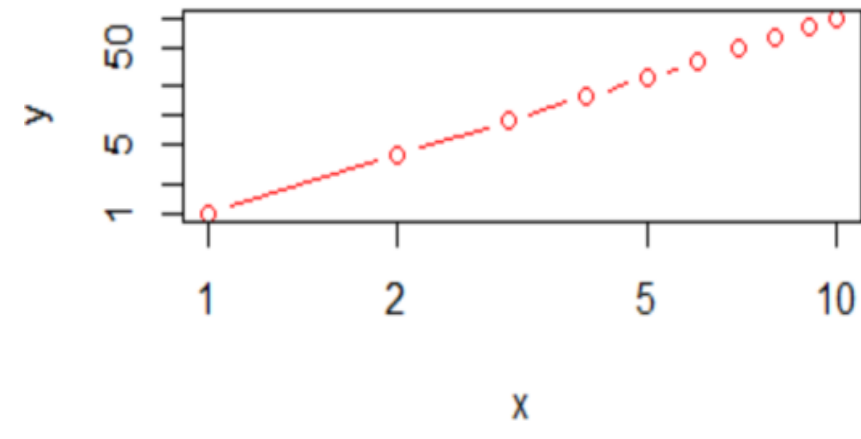
```
plot(  
  1:10,  
  Power3(1:10,2),  
  log = "y",  
  main = 'y = x^2, log="y"',  
  xlab = "x",  
  ylab = "y",  
  type = "b",  
  col = "red"  
)
```

$y = x^2, \text{log}="y"$



```
plot(  
  1:10,  
  Power3(1:10,2),  
  log = "xy",  
  main = 'y = x^2, log="xy"',  
  xlab = "x",  
  ylab = "y",  
  type = "b",  
  col = "red"  
)
```

$y = x^2, \text{log}="xy"$



Group Assignment: Problem 12 – Function

(f) PlotPower(x,a): plot $y = x^a$

Call PlotPower(1:10, 3)

```
PlotPower=function(x,a){  
  plot(  
    x,  
    x^a,  
    main = "y = x^a",  
    xlab = "x",  
    ylab = "y"  
  )  
}
```

$y = x^a$

