STAT435 Intro to Statistical Machine Learning

Week 1: Logistics, R, and R Markdown

Richard Li

Mar 29

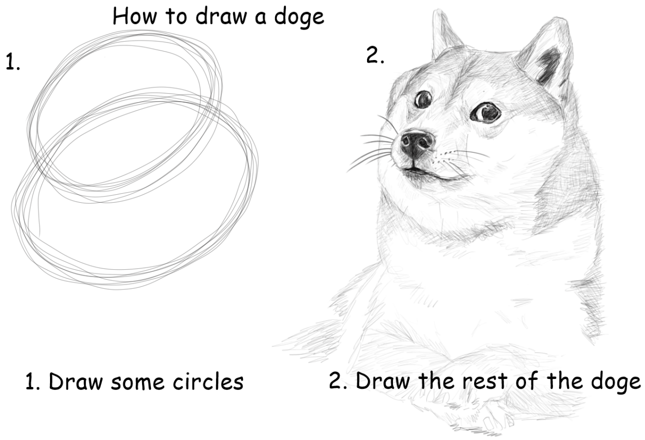
About me

- Zehang Li, go by Richard
- Office: PDL C-14G
    - *Looking for a big metal door at LL level of PDL C wing*
- Office hour: Thursday 2:30 - 3:30
    - *Remember homeworks are due on Fridays :)*

About the course

Machine learning in a nutshell

- Understand data
- Do statistics
- Write codes



Week 1 quick pool

Wednesday sessions

- Highlight/Review/Supplement lectures?
- R programming demo?
- Additional exercise problems?
- ~~Kayaking, Barbecue, Deriving homework problems~~

What we'll do today

Introduction to R

- Haven't used R at all?
  - *Don't worry, it's just another programming language*
- Haven't programmed at all?
  - *Don't worry, it's just a fancy calculator (sort of)*

Documentation with R

- R Markdown

Start using R

R

- a programming language and software environment for statistics
- many packages to use for even very complicated models
- Very easy setup with most OS environments (most of the time)
- Download: https://cloud.r-project.org

RStudio

- Download: http://www.rstudio.com/download
- Using Rstudio is totally optional, but usually it makes life easier
- Syntax highlighting
- Nice organization of windows
- Auto-saving codes when crashes
- Much easier for R Markdown

Try for yourself

```
print("Hello, World!")
```

Basics

- Highlight codes in the editor window and click Run or hit Cntl-Enter (Command-Enter on a Mac) to run
- Type lines in the console and press Enter
- Making sure the the lines you typed are finished
  - *If not, you will see a '+' in front of the line*
  - *Finish the line or hit ESC to escape*
- Now, try calculating "8 + 24 + 23" and $\sqrt{2}$ in R
- Try look for help with functions using '?'

Example from ISL

- Section 2.3 from ISL
  - *Construct vectors, matrices, and draw random numbers*
  - *Basic scatter plot and contour plots*
  - *Load ans summarize data*
- Section 3.6 from ISL
  - *Simple linear regression*
  - *Multiple regression*
- Section 4.6.5 from ISL
  - *KNN (homework)*

Basic R codes

Initialize and construct vectors

```
x <- c(1, 2, 3, 4, 6)
x
```

```
## [1] 1 2 3 4 6
```

```
x <- c(6:1)
x
```

```
## [1] 6 5 4 3 2 1
```

```
length(x)
```

```
## [1] 6
```

```
y <- matrix(x, nrow = 2, ncol = 3)
y
```

```
##      [,1] [,2] [,3]
## [1,]    6    4    2
## [2,]    5    3    1
```

```
y <- matrix(x, nrow = 2, ncol = 3, byrow=TRUE)
y
```

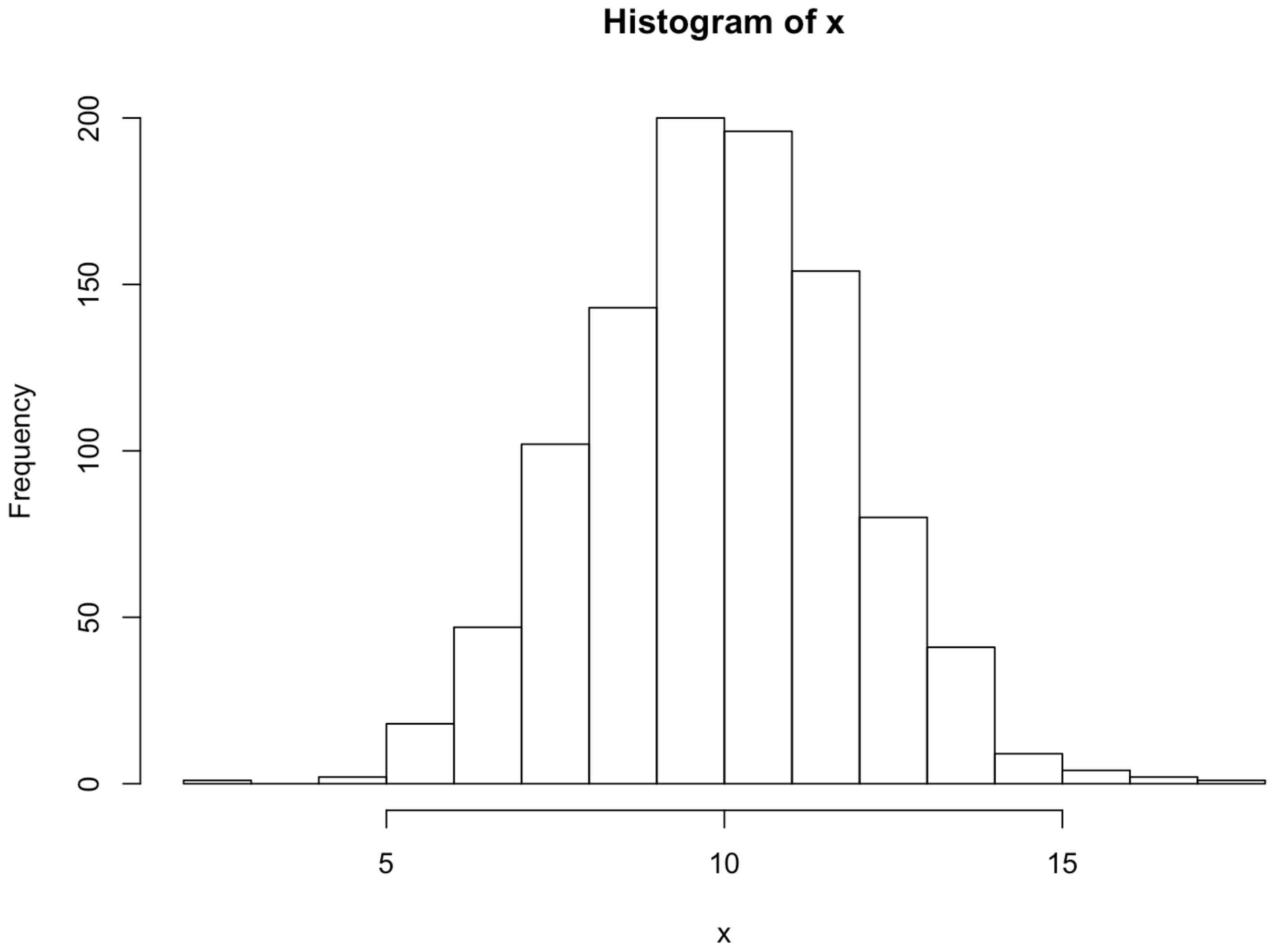
```
##      [,1] [,2] [,3]
## [1,]    6    5    4
## [2,]    3    2    1
```

Generate Random numbers following normal distribution

```
x <- rnorm(1000, mean = 10, sd = 2)
head(x)
```

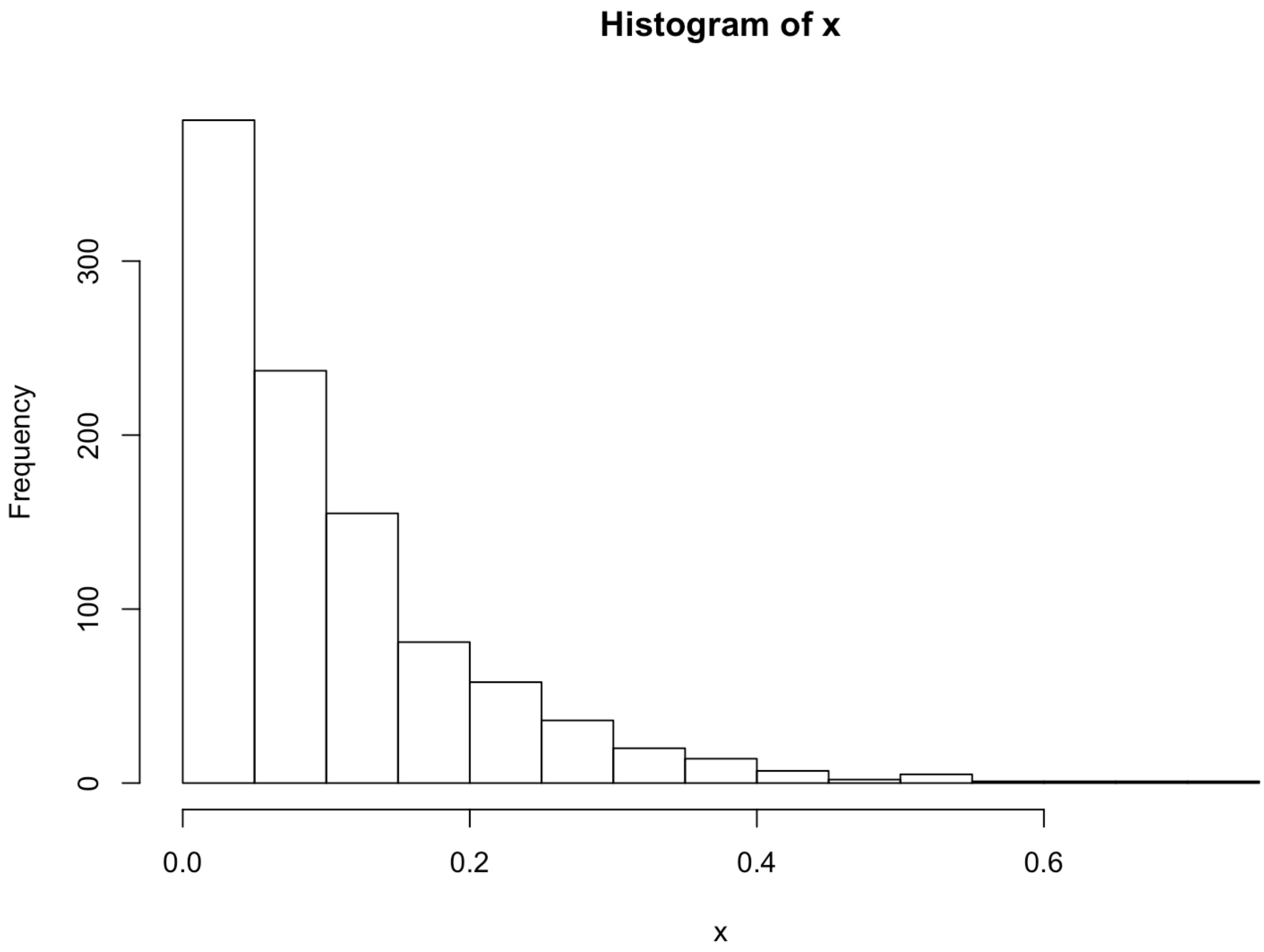
```
## [1]  7.658101  9.815746  9.594886 11.593327  9.968943  7.758255
```

```
hist(x)
```



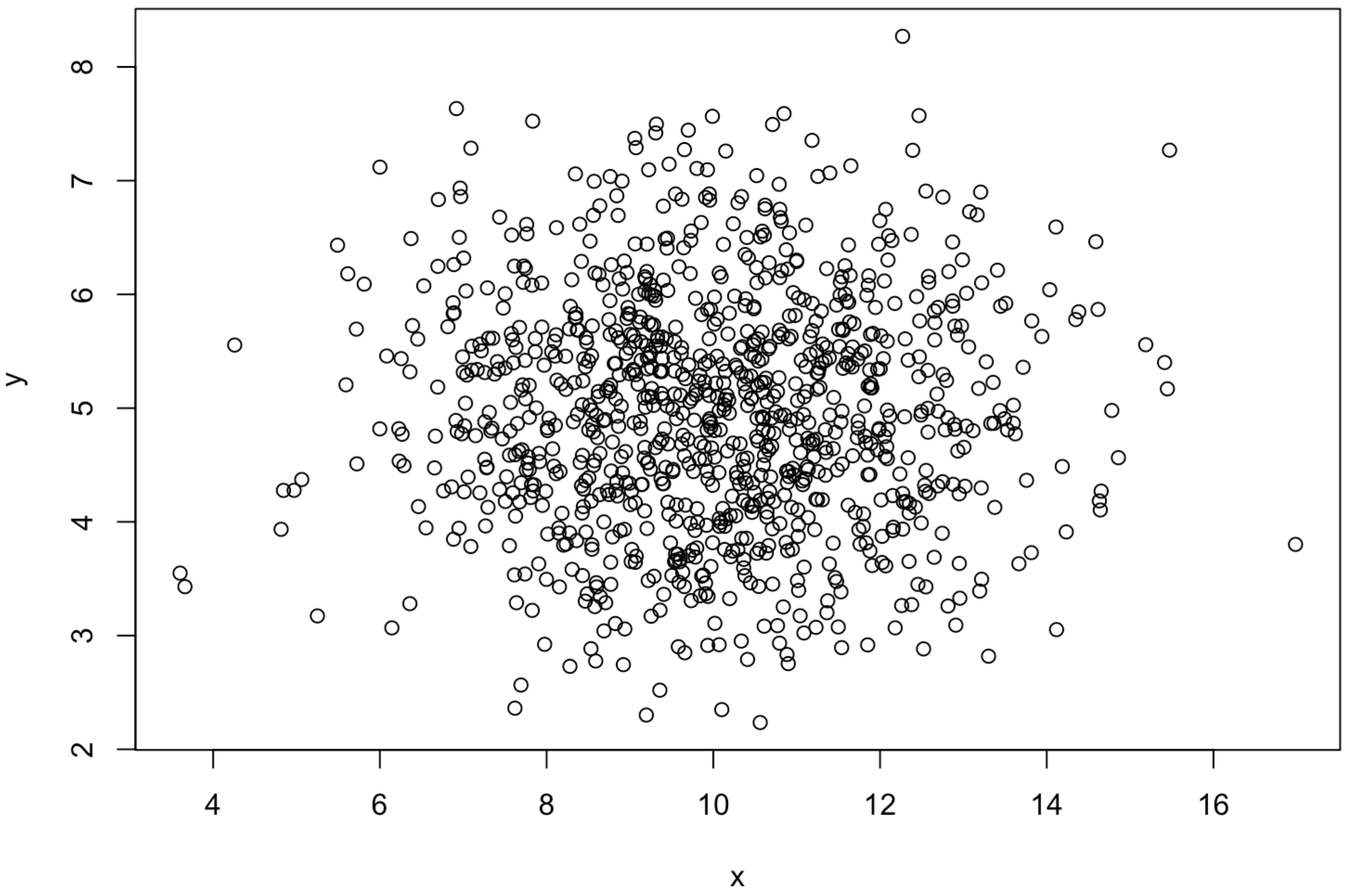
Generate Random numbers following other distributions

```r
x <- rexp(1000, rate = 10)
hist(x)
```



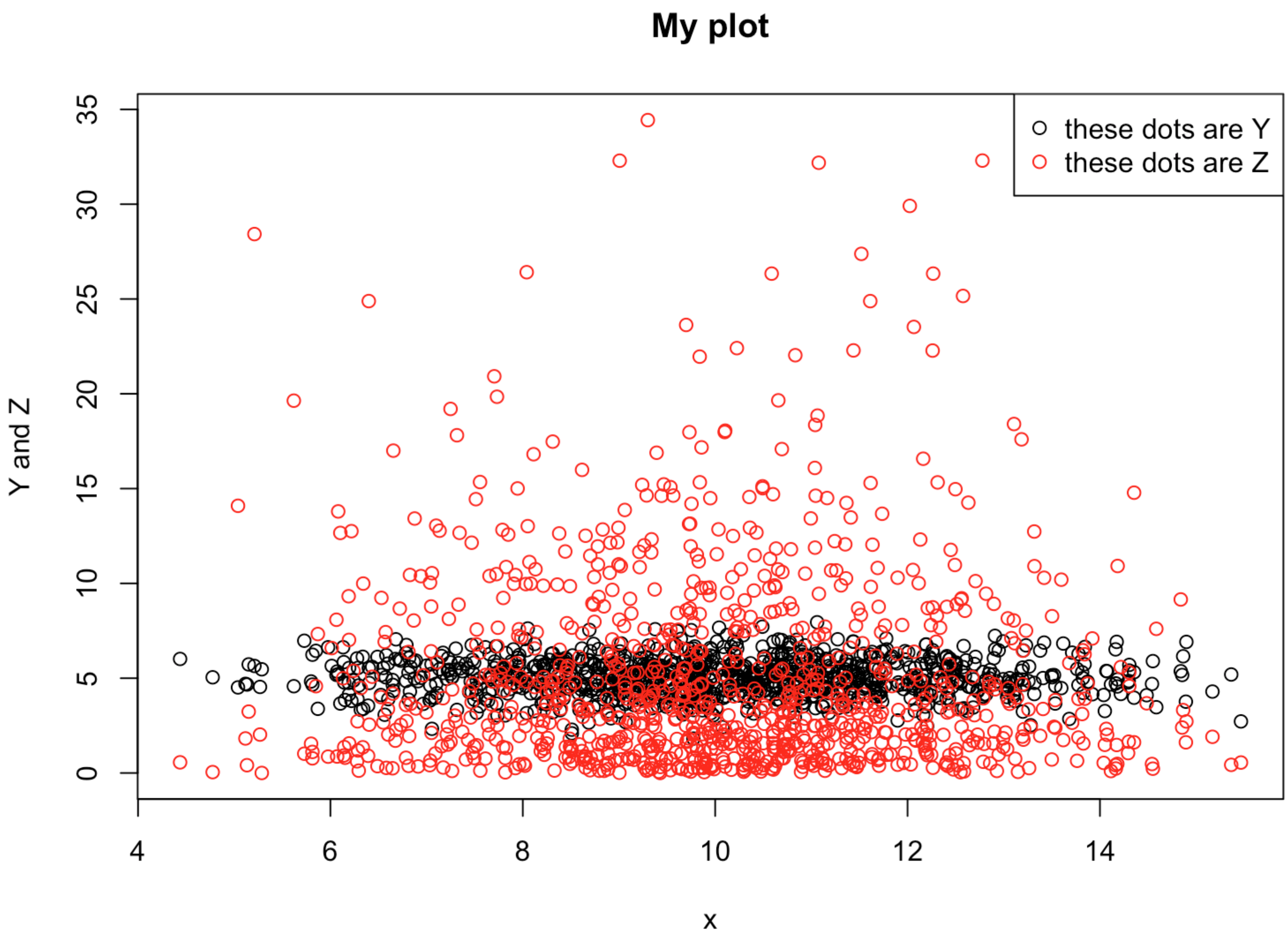
Scatter plot

```r
x <- rnorm(1000, mean = 10, sd = 2)
y <- rnorm(1000, mean = 5, sd = 1)
plot(x, y)
```



Customizing scatter plots

```r
x <- rnorm(1000, mean = 10, sd = 2)
y <- rnorm(1000, mean = 5, sd = 1)
z <- rexp(1000, rate = 0.2)
plot(x, y, ylim = range(c(x, y, z)), main = "My plot", ylab="Y and Z")
points(x, z, col = "red")
legend("topright", c("these dots are Y", "these dots are Z"),
      pch = c(1, 1), col = c("black", "red"))
```


My plot
Y and Z
0
5
10
15
20
25
30
35
x
4
6
8
10
12
14
○ these dots are Y
○ these dots are Z

Regression example

Read data

- The MASS library contains the Boston data set, which records *medv* (median house value) for 506 neighborhoods around Boston.
- Predict *medv* using 13 predictors such as
  - *rm (average number of rooms per house),*
  - *age (average age of houses),*
  - *lstat (percent of households with low socioeconomic status).*

```
library(MASS)
data(Boston)
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"
##  [9] "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

Regression

- You should have seen this before

```
lm.fit <- lm(medv ~ lstat + age, data=Boston)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = medv ~ lstat + age, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.981  -3.978  -1.283   1.968  23.158
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.22276    0.73085  45.458  < 2e-16 ***
## lstat       -1.03207    0.04819 -21.416  < 2e-16 ***
## age          0.03454    0.01223   2.826  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.173 on 503 degrees of freedom
## Multiple R-squared:  0.5513, Adjusted R-squared:  0.5495
## F-statistic:   309 on 2 and 503 DF,  p-value: < 2.2e-16
```

Diagnostics

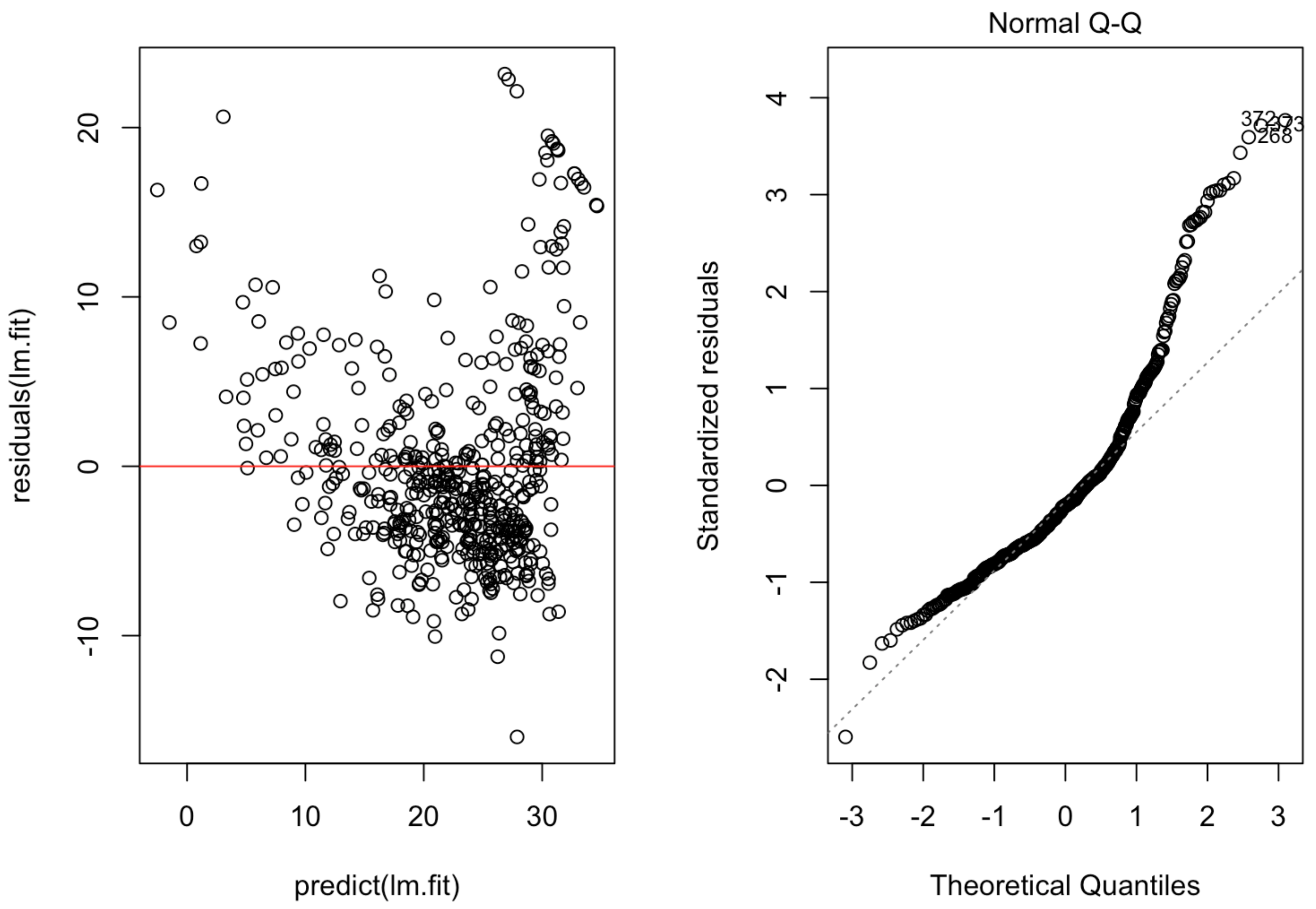
- How about confidence intervals for the regression coefficients

```
confint(lm.fit)
```

```
##                   2.5 %      97.5 %
## (Intercept) 31.78687150 34.65864956
## lstat       -1.12674848 -0.93738865
## age          0.01052507  0.05856361
```

- Visual check of
  - *residual against fitted values*
  - *QQ plot of residuals*
- *Why do we want to see these plots?*

```
par(mfrow = c(1, 2))
plot(predict(lm.fit), residuals(lm.fit))
abline(h=0, col = "red")
plot(lm.fit, which = 2)
```



More regressors

```
lm.fit.all <- lm(medv ~ ., data=Boston)
```

- *Read more in book!*

Trick of R programming

Practice!

Fun stuff, finally, what we are all here for :)

```r
install.packages("fun")
library(fun)
gomoku(n = 19)
```

Try on your laptop :) Then we’ll talk some serious business.

R Markdown

- Reporting tool to combine everything together
  - *code*
  - *result*
  - *comment/discussion*
- Easy to get started with RStudio
- Good for Homework
  - *Not required though, MS Word is perfectly accept as long as it is clear and readable*

R Markdown Demo

- Choose File > New File > R Markdown…
- Make sure HTML output is selected and click OK
- Save the file somewhere, call it demo.Rmd
- Click the Knit HTML button
- Find the HTML file and open in browser

R Markdown Demo

- *Saving a PDF file is tricky*
- You will need to install $\TeX$ on your computer
  - *https://www.latex-project.org/get/*
- An Alternative way is to print your HTML file to PDF
- *For HW submission*, HTML file is enough. But to be absolutely safe, upload both HTML file and RMD file.

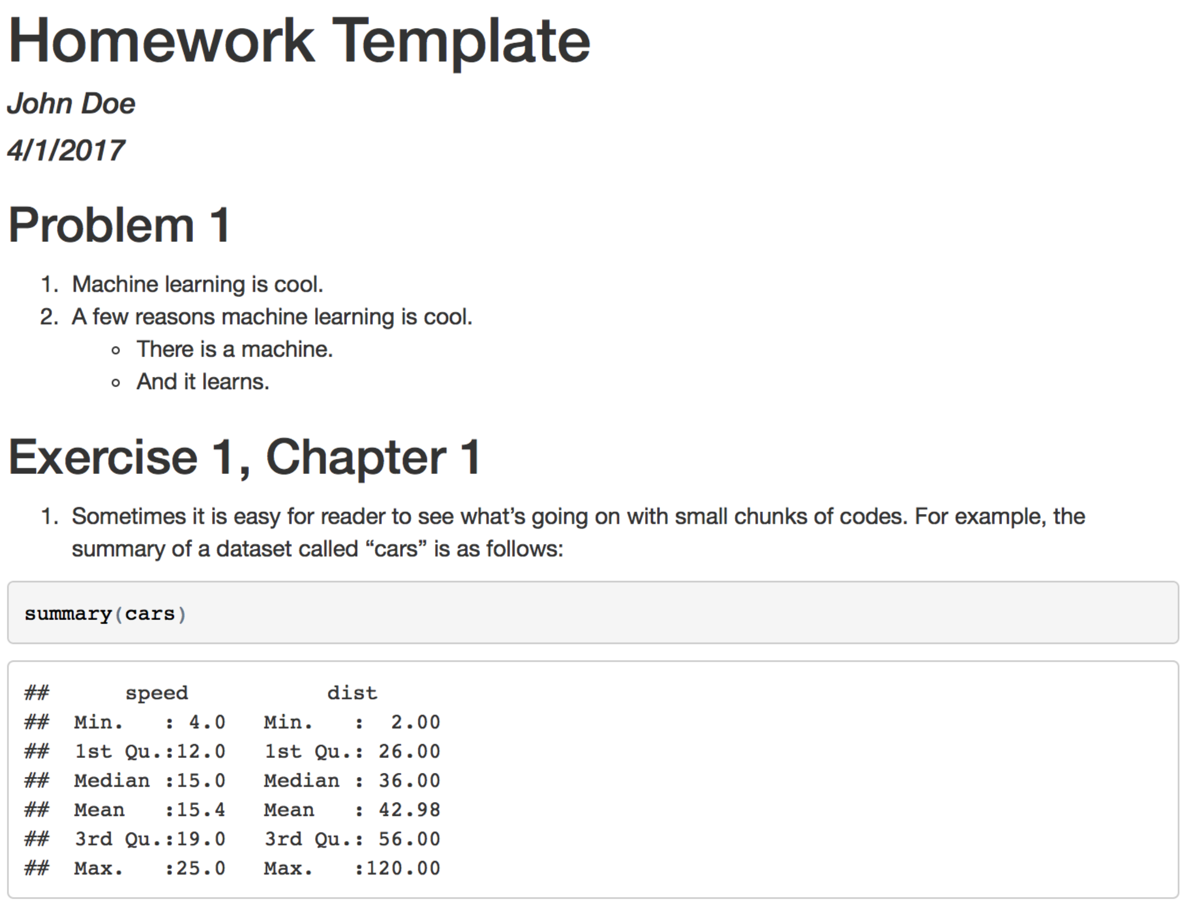
R Markdown Syntax

- Header block
- **bold**: `**bold**`
- *italic*: `*italic*` or `_italic_`
- Header: `# Header`
- Subheader: `## Subheader`
- Subsubheader: `### Subsubheader`
- Code chunks:

~~~
```{r}
x <- 1:10
y <- 2:11
plot(x, y)
```
~~~

R Markdown for Homework

Check out the Homework template on canvas!



Homework Template

John Doe

4/1/2017

Problem 1

1. Machine learning is cool.
2. A few reasons machine learning is cool.
   - There is a machine.
   - And it learns.

Exercise 1, Chapter 1

1. Sometimes it is easy for reader to see what's going on with small chunks of codes. For example, the summary of a dataset called "cars" is as follows:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

More logistics, class Resources

Office hours or email me to schedule a time

Canvas discussion board

- Phrase your question so that other people can answer
- Give codes that other people can run *and replicate your problem*

Bad examples

1. I ran the regression codes but it didn't work
2. I have ``lm(y ~ x)" in my codes but it didn't work

Good example

I tried the following codes and the error message says: variable lengths differ (found for 'x')

```
x <- seq(1:10)
y <- seq(1:100)
model <- lm(y ~ x)
```

External Resources

- R for Data Science book
  - *Everything about R with no stones unturned*
- Lecture materials from CSSS 508: Introduction to R for Social Scientists
  - *Advanced materials on data structure, fancy plots, etc.*
- *Stack Overflow!*

Questions?