Week 4: Linear Correlation and Regression

Example 1

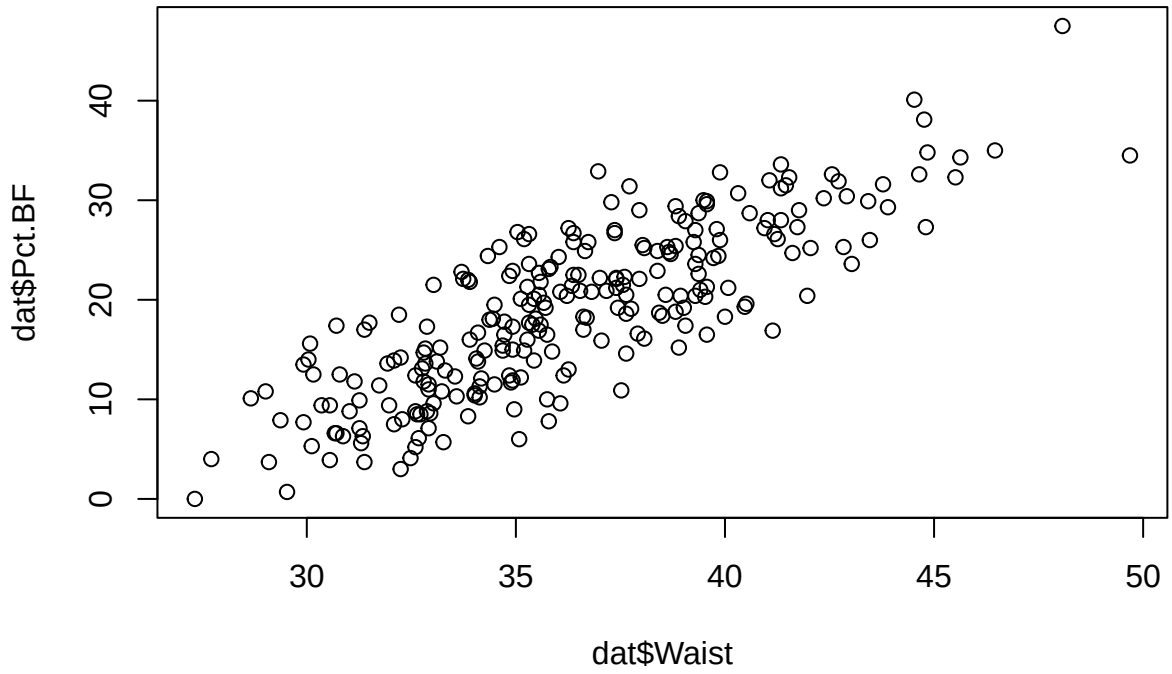
A study of men's health measured 14 body characteristics of 250 men. We import the data from Professor Nick Horton's website.

```
dat <- read.csv(url("https://nhorton.people.amherst.edu/is5/data/Bodyfat.csv"))
head(dat)
```

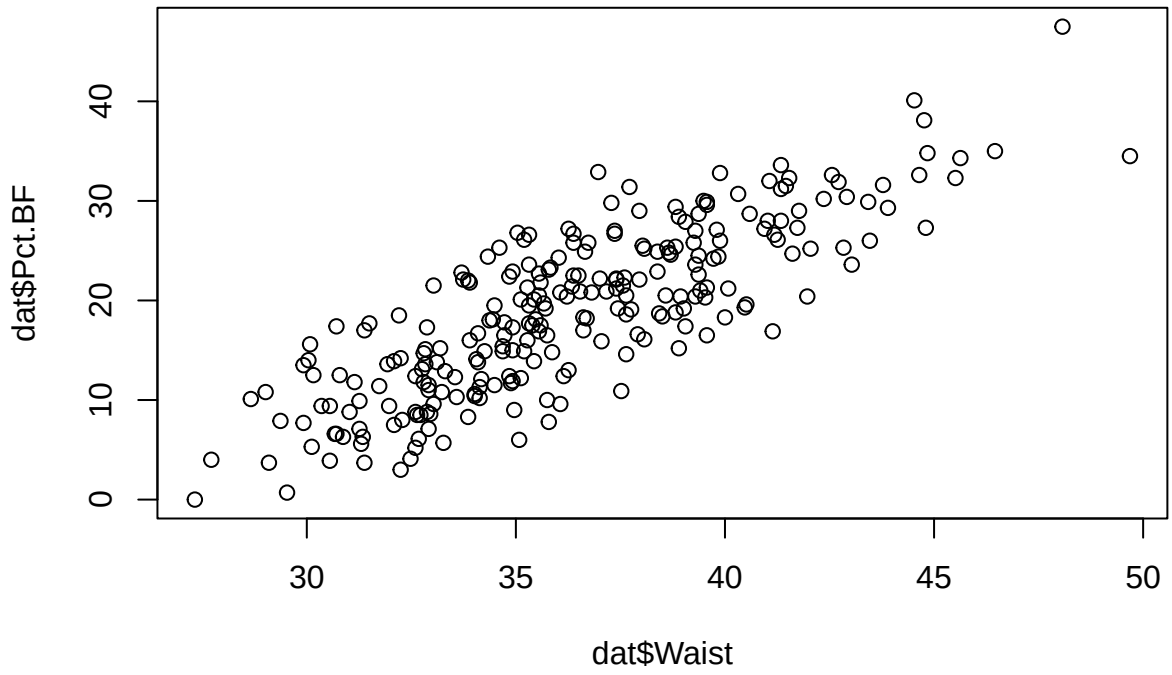
```
##   Density Pct.BF Age Weight Height Neck Chest Abdomen    Waist   Hip Thigh Knee
## 1  1.0708   12.3  23 154.25  67.75 36.2  93.1    85.2 33.54331  94.5  59.0 37.3
## 2  1.0853    6.1  22 173.25  72.25 38.5  93.6    83.0 32.67717  98.7  58.7 37.3
## 3  1.0414   25.3  22 154.00  66.25 34.0  95.8    87.9 34.60630  99.2  59.6 38.9
## 4  1.0751   10.4  26 184.75  72.25 37.4 101.8    86.4 34.01575 101.2  60.1 37.3
## 5  1.0340   28.7  24 184.25  71.25 34.4  97.3   100.0 39.37008 101.9  63.2 42.2
## 6  1.0502   20.9  24 210.25  74.75 39.0 104.5    94.4 37.16535 107.8  66.0 42.0
##   Ankle Bicep Forearm Wrist
## 1  21.9  32.0    27.4  17.1
## 2  23.4  30.5    28.9  18.2
## 3  24.0  28.8    25.2  16.6
## 4  22.8  32.4    29.4  18.2
## 5  24.0  32.2    27.7  17.7
## 6  25.6  35.7    30.6  18.8
```

To make a scatter plot of body fat percentage and waist size, we can do either of the following.

```
plot(dat$Waist, dat$Pct.BF)
```




```
plot(dat$Pct.BF ~ dat$Waist)
```



To find the least squares line, we use the lm function.

```
fit <- lm(dat$Pct.BF ~ dat$Waist)
summary(fit)
```

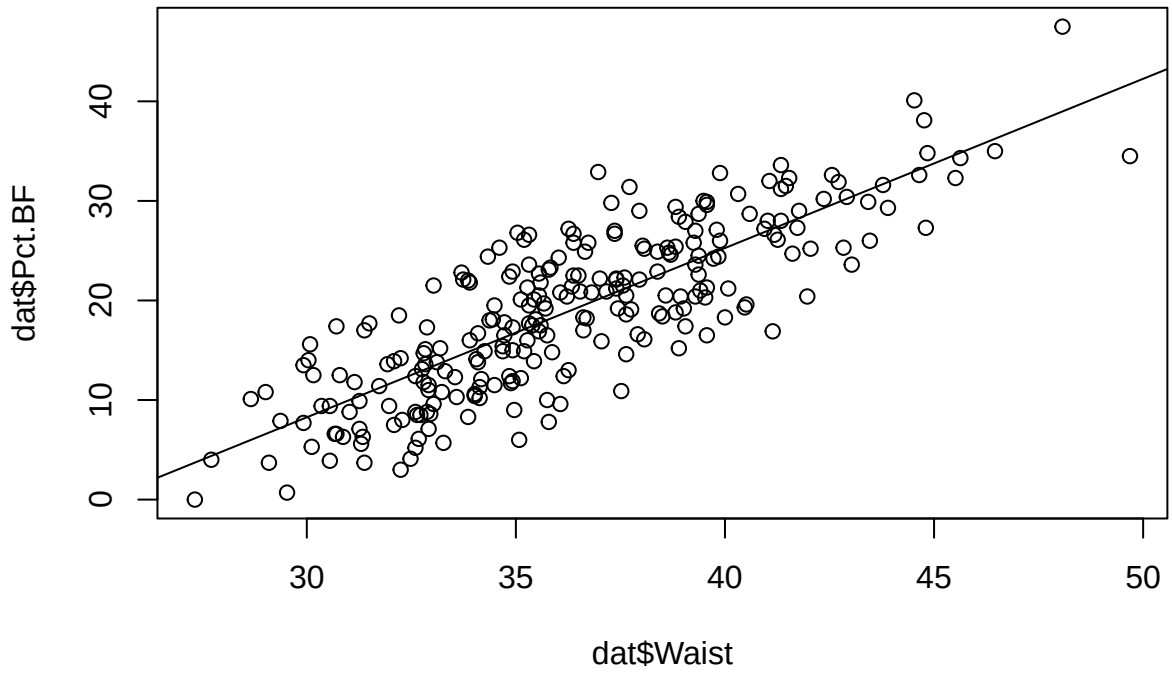
```
##
## Call:
## lm(formula = dat$Pct.BF ~ dat$Waist)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8987  -3.6453   0.1864   3.1775  12.7887
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.73413    2.71651  -15.73   <2e-16 ***
## dat$Waist     1.69997    0.07431   22.88   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.713 on 248 degrees of freedom
## Multiple R-squared:  0.6785, Adjusted R-squared:  0.6772
## F-statistic: 523.3 on 1 and 248 DF,  p-value: < 2.2e-16
```

From the R output we read the equation

$$\hat{y} = 1.70x - 42.73$$

where x is the waist size and R^2 value 0.6785. We add the least squares line to the scatter plot.

```
plot(dat$Pct.BF ~ dat$Waist)
abline(fit)
```



Example 2

A survey was conducted in the United States and 10 countries of Western Europe to determine the percentage of teenagers who had used marijuana and other drugs. The results are below.

```
dat <- read.csv(url("https://nhorton.people.amherst.edu/is5/data/Drug_abuse.csv"))
dat
```

```
##       Country Marijuana.... Other.Drugs....
## 1    CzechRep            22               4
## 2     Denmark            17               3
## 3     England            40              21
## 4     Finland             5               1
## 5     Ireland            37              16
## 6       Italy            19               8
## 7  No.Ireland            23              14
## 8      Norway             6               3
## 9    Portugal             7               3
## 10   Scotland            53              31
## 11        USA            34              24
```

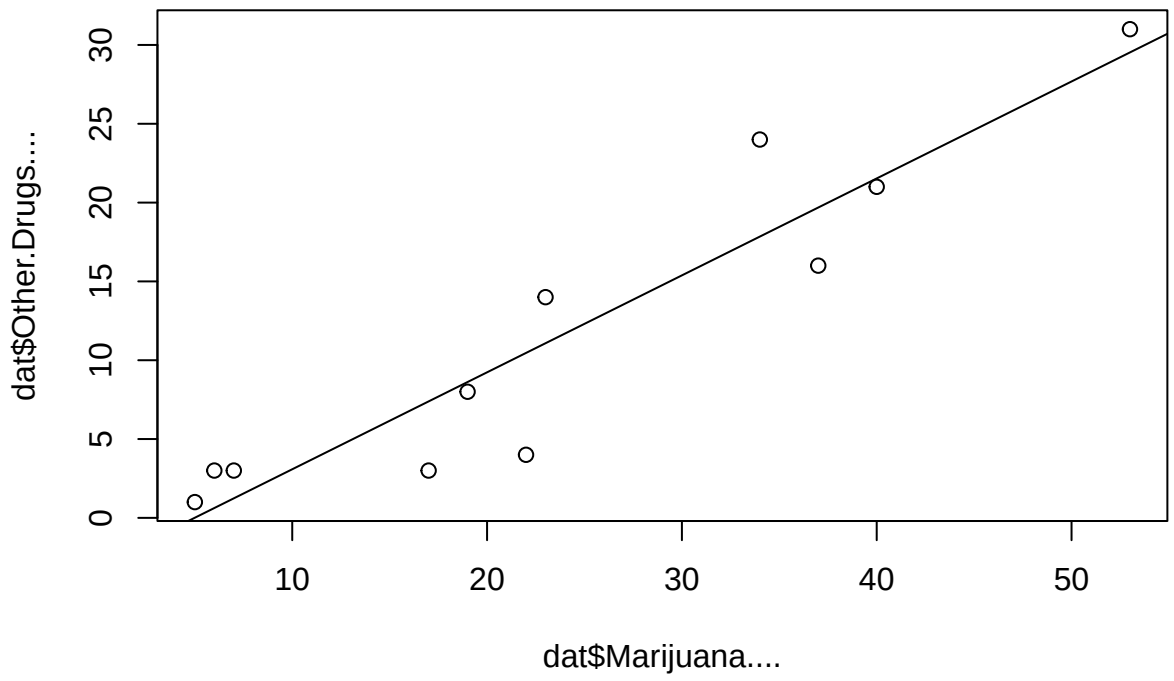
We create a scatter plot, and find the least square line.

```
plot(dat$Other.Drugs.... ~ dat$Marijuana....)
fit <- lm(dat$Other.Drugs.... ~ dat$Marijuana....)
summary(fit)
```

```
##
## Call:
## lm(formula = dat$Other.Drugs.... ~ dat$Marijuana....)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4623 -2.1523  0.9928  2.0703  6.1577
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -3.06780    2.20436  -1.392    0.197
## dat$Marijuana....  0.61500    0.07835   7.849 2.58e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.853 on 9 degrees of freedom
## Multiple R-squared:  0.8725, Adjusted R-squared:  0.8584
## F-statistic: 61.61 on 1 and 9 DF,  p-value: 2.576e-05
```

```
abline(fit)
```



Follow up

How are fat and protein related on the entire Burger King's menu? We import the data set as shown below.

```
dat <- read.csv(url("https://nhorton.people.amherst.edu/is5/data/Burger_King_items.csv"))
names(dat)
```

```
##  [1] "Item"          "Serving.size"  "Calories"      "Fat.Cal"
##  [5] "Protein.g."    "Fat.g."        "Sat.Fat.g."    "Trans.fat.g."
##  [9] "Chol.mg."      "Sodium.mg."    "Carbs.g."      "Fiber.g."
## [13] "Sugar.g."      "Meat"          "Breakfast"     "Not.Breakfast"
## [17] "CarbsxMeat"
```

Use the methods demonstrated above to make a scatter plot and find the least squares line. Hint: See Wang page 45.