Lab 1

Akanksha

OpenIntro Biostatistics

Topics

- Dataset manipulation in R
- Numerical summaries: mean, SD, median, IQR
- Graphical summaries: boxplots, histograms, scatterplots

The first two sections of this lab introduce basic tools for working with data matrices, as well as the commands for producing numerical and graphical summaries. The last section focuses on data interpretation and reinforces the statistical concepts presented in the text. The material in this lab corresponds to Sections 1.1 - 1.2 and 1.4 - 1.6 of *OpenIntro Biostatistics*.

Section 1: BRFSS.

The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. The BRFSS is designed to identify risk factors in the adult population and report emerging health trends. For example, respondents are asked about their diet, weekly exercise, possible tobacco use, and healthcare coverage.

1. Use the following command to download the dataset `cdc` from a URL. This dataset is a sample of 20,000 people from the survey conducted in 2000, and contains responses from a subset of the questions asked on the survey.

```
source("http://www.openintro.org/stat/data/cdc.R")
```

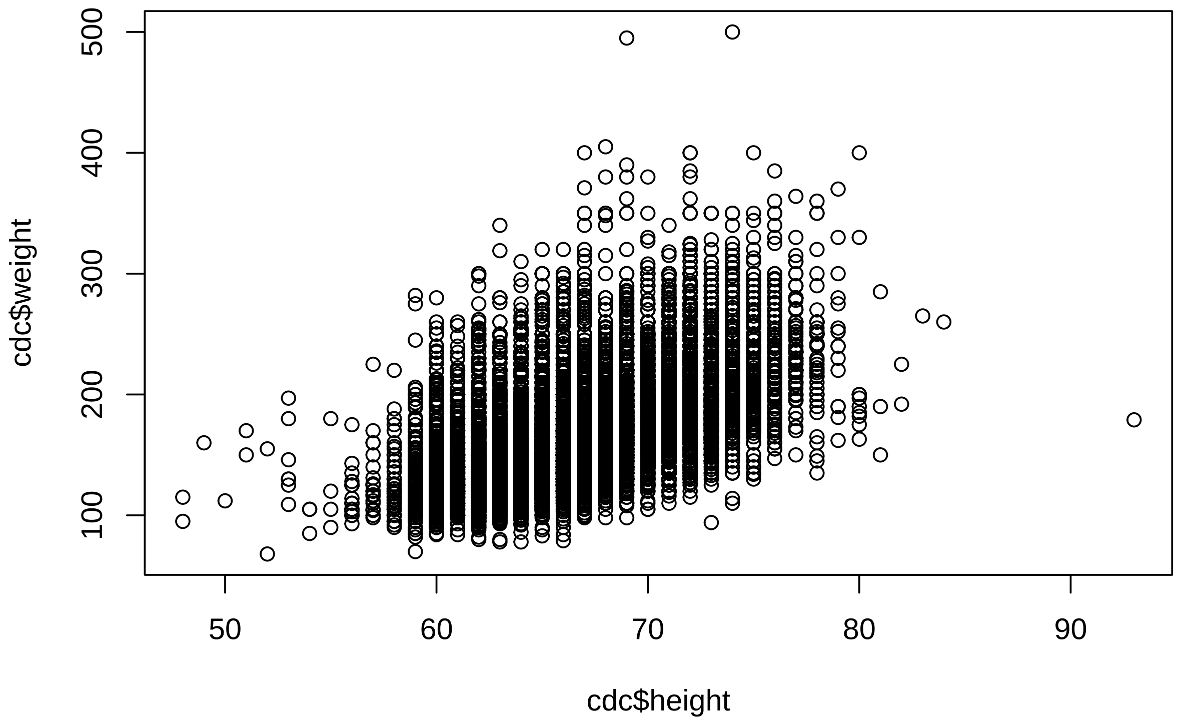
2. Take a look at the Environment tab, where `cdc` should now be visible. Click the blue button next to the dataset name to view a summary of the 9 variables contained in the data matrix. To view the dataset itself, click on the name of the dataset; alternatively, enter the command

```
View(cdc)
```

```
Each row of the data matrix represents a case and each column represents a variable. Each vari
```

3. The $ operator in R is used to access variables within a dataset; for example, `cdc$height` tells R to look in the `cdc` dataframe for the `height` variable. Make a scatterplot of `height` and `weight` using the `plot( )` command:

```r
plot(cdc$weight ~ cdc$height)
```



```
Do \texttt{height} and \texttt{weight} appear to be associated?

\textcolor{NavyBlue}{The visible upward trend in the cloud of points shows that \texttt{height
```

4. The conversion from inches to meters is 1 in = .0254 m. Create a new variable `height.m` that records height in meters. Similarly, the conversion from pounds to kilograms is 1 lb = .454 kg. Create a new variable `weight.kg` that records weight in kilograms.

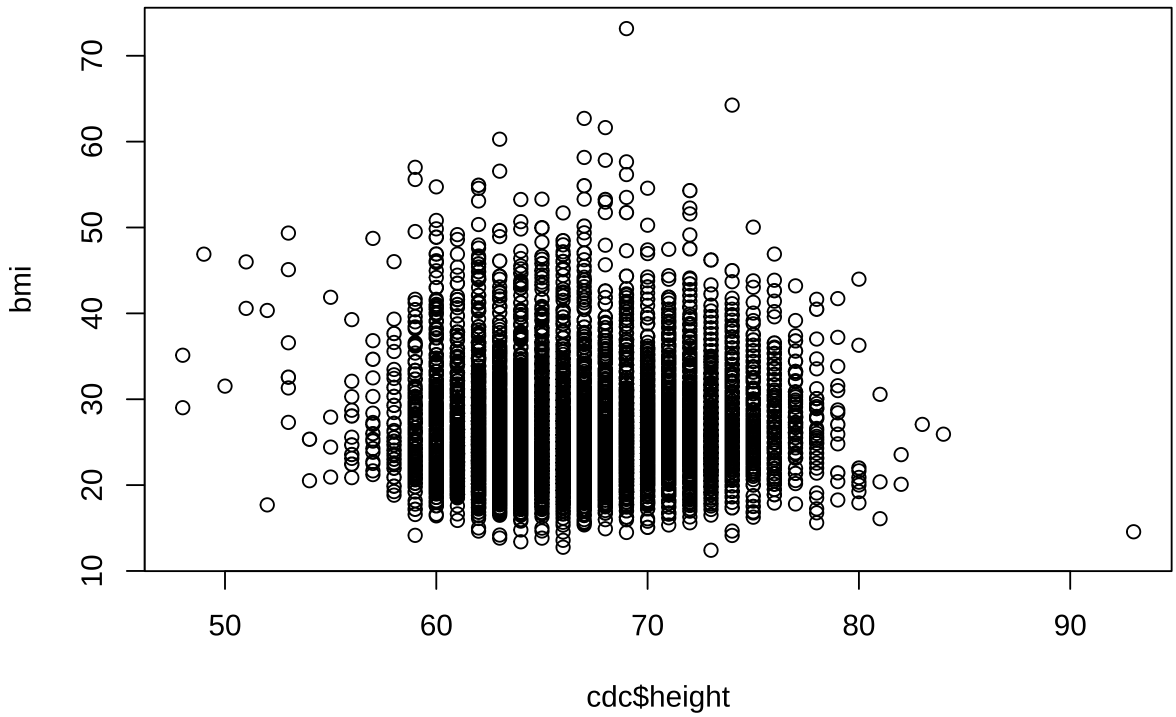
```r
#create height.m
height.m = cdc$height*.0254

#create weight.kg
weight.kg = cdc$weight*.454
```

5. BMI is calculated as weight in kilograms divided by height squared. Create a new variable `bmi` and make a scatterplot of `height` and `BMI`. Do `height` and `BMI` seem to be associated?

```
#create bmi
bmi = (weight.kg)/(height.m^2)

#plot height and bmi
plot(cdc$height, bmi)
```



```
\textcolor{NavyBlue}{Height and BMI do not appear to be associated.}
```

```
A BMI of 30 or above is considered obese. Why might health agencies choose to use BMI as a mea
```

```
\textcolor{NavyBlue}{Since \texttt{height} and \texttt{BMI} have a much weaker association, it
```

6. Row-and-column notation in combination with square brackets can be used to access a subset of the data. For example, to access the sixth variable (`weight`) of the 567th respondent, use the command:

```
cdc[567, 6]
```

```
## [1] 160
```

To see the weight for the first ten respondents, use:

```
cdc[1:10, 6]
```

```
##  [1] 175 125 105 132 150 114 194 170 150 180
```

If the column number is omitted, then all the columns will be returned for rows 1 through 10:

```
cdc[1:10, ]
```

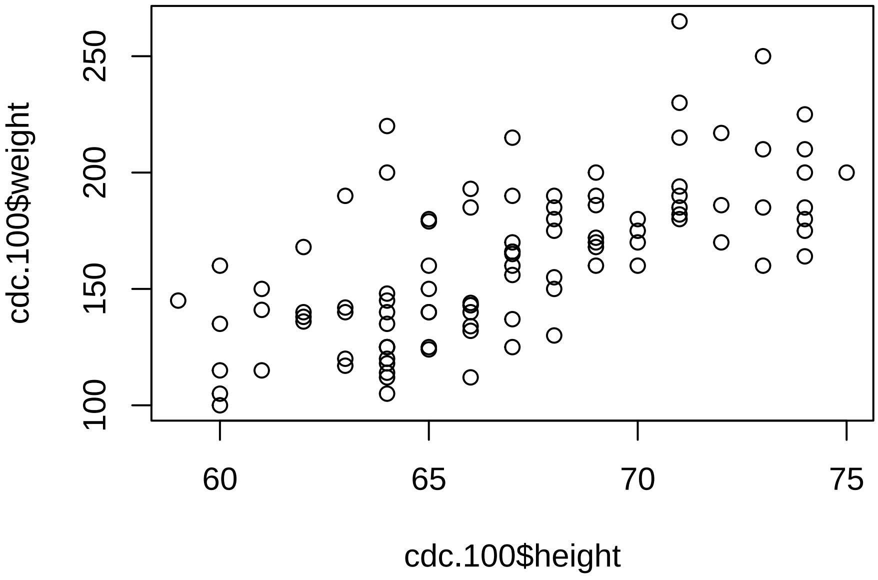
```
##      genhlth exerany hlthplan smoke100 height weight wtdesire age gender
## 1       good       0        1        0     70    175      175  77      m
## 2       good       0        1        1     64    125      115  33      f
## 3       good       1        1        1     60    105      105  49      f
## 4       good       1        1        0     66    132      124  42      f
## 5  very good       0        1        0     61    150      130  55      f
## 6  very good       1        1        0     64    114      114  55      f
## 7  very good       1        1        0     71    194      185  31      m
## 8  very good       0        1        0     67    170      160  45      m
## 9       good       0        1        1     65    150      130  27      f
## 10      good       1        1        0     70    180      170  44      m
```

Likewise, omit the range for the rows to access all observations for column 6. The following wi

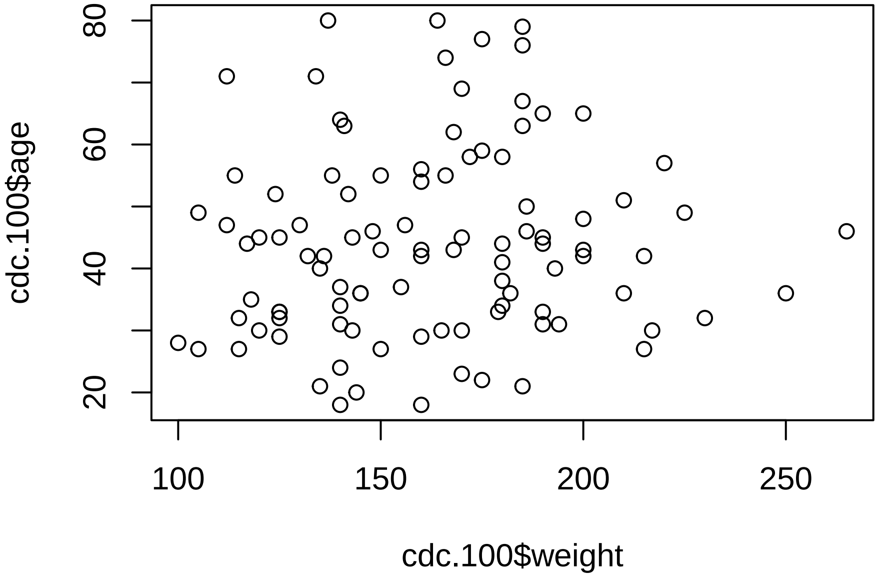
```
cdc[ ,6] #results of this chunk are hidden with eval = FALSE
```

7. Use bracket notation to make a scatterplot of `height` and `weight` for the first 100 respondents. There are multiple ways to do this—find one that works!

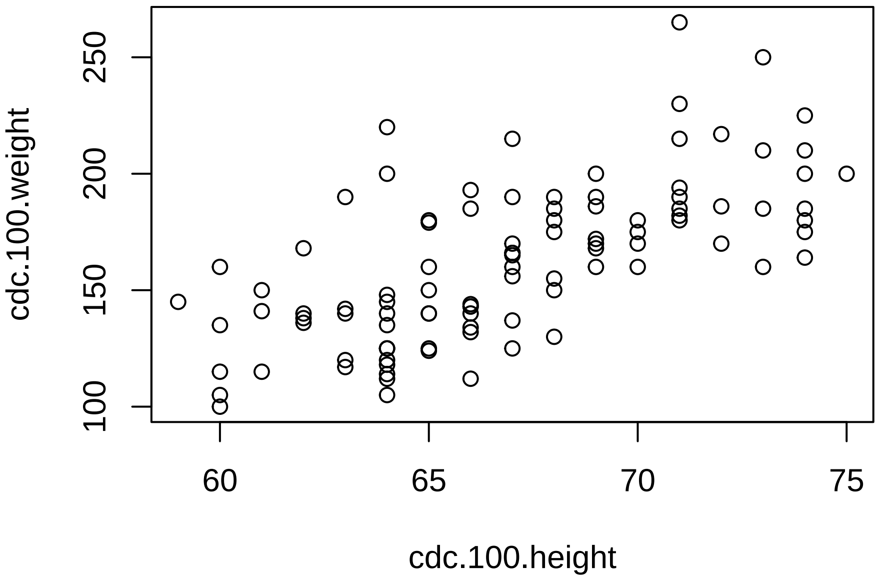
```
#create a new dataset with just 100 observations
cdc.100 = cdc[1:100, ]
plot(cdc.100$height, cdc.100$weight)
```



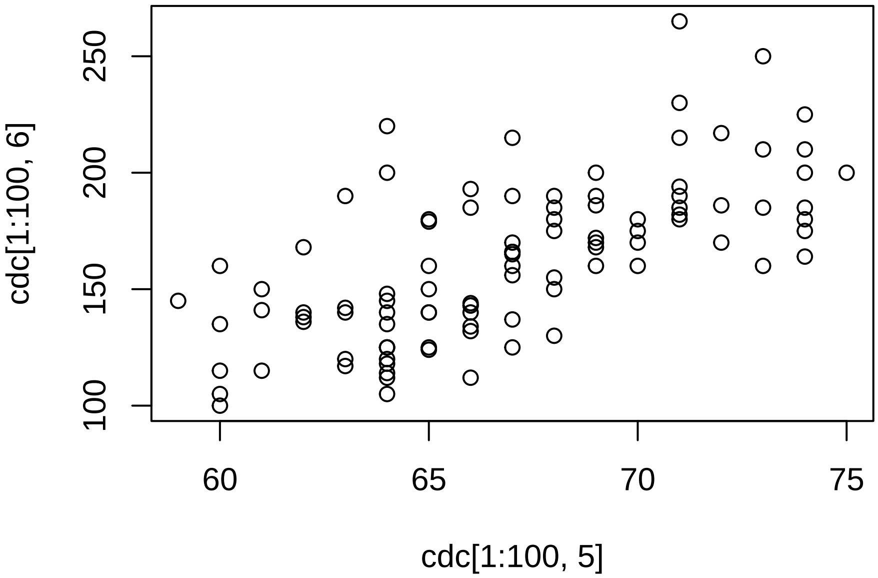
```
plot(cdc.100$weight, cdc.100$age)
```



```
#subset the variables separately
cdc.100.weight = cdc[1:100, 6]
cdc.100.height = cdc[1:100, 5]
plot(cdc.100.height, cdc.100.weight)
```



```
#nest the commands
plot(cdc[1:100, 5], cdc[1:100, 6])
```


cdc[1:100, 6]
100
150
200
250
60
65
70
75
cdc[1:100, 5]