Dunn book reading group Chapter 01

Ed Harris

12/02/2020

Dunn Ch01 notes

Preface

- Book is intended for wide audience
    - (“Second stats course” up to more advanced)
- Regression approach
- Build on regression towards Generalized Linear Model
- {GLMsData} package

1.2 Data (and conventions)

lungcap data

lungcap {GLMsData} The health and smoking habits of 654 youth

Age the age of the subject in completed years; a numeric vector

FEV the forced expiratory volume in litres, a measure of lung capacity; a numeric vector

Ht the height in inches; a numeric vector

Gender the gender of the subjects: a numeric vector with females coded as 0 and males as 1

Smoke the smoking status of the subject: a numeric vector with non-smokers coded as 0 and smokers as 1

```
library(GLMsData)  #Load the {GLMsData} package
data(lungcap)      #Load lungcap data into memory
dim(lungcap)       #Dimensions (rows and columns of data)
```

```
## [1] 654   5
```

```
head(lungcap)      #Print first 6 rows
```

```
##   Age   FEV Ht Gender Smoke
## 1   3 1.072 46      F     0
## 2   4 0.839 48      F     0
## 3   4 1.102 48      F     0
## 4   4 1.389 48      F     0
## 5   4 1.577 49      F     0
## 6   4 1.418 49      F     0
```

Notes on notation

$y_i = x_{1i} + x_{2i}$

equivalent to
dependent_var = first_ind_var + Second_ind_var

equivalent to
FEV = Age + Ht

NB variable addressing notation
y_i is the ith observation of variable y

e.g. y_3
equivalent to

```
lungcap$FEV[3]
```

```
## [1] 1.102
```

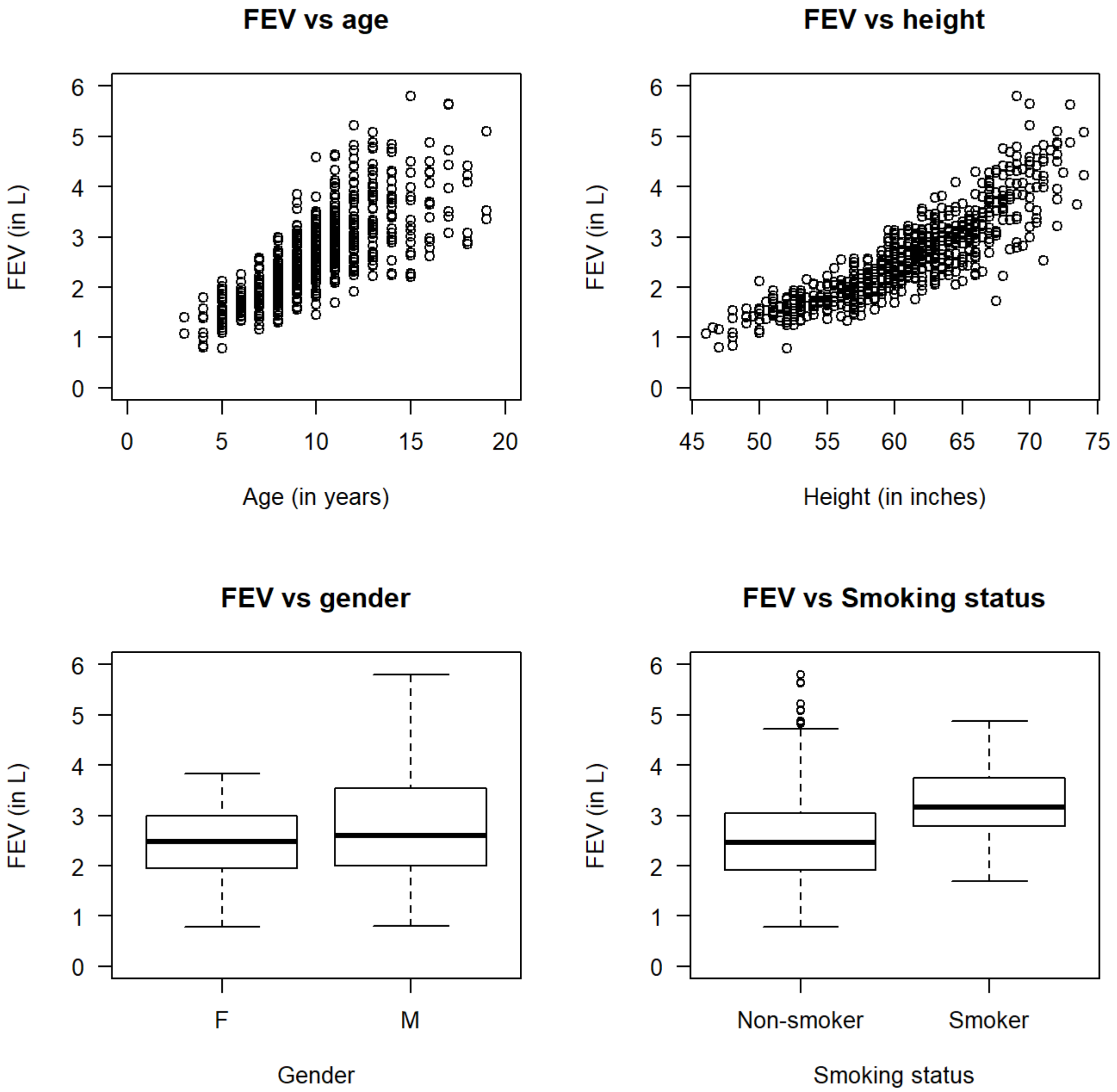
Factor issues with variable Smoke

```
lungcap$Smoke <- factor(lungcap$Smoke, levels=c(0, 1), # The values of Smoke
labels=c("Non-smoker","Smoker"))
table(lungcap$Smoke)
```

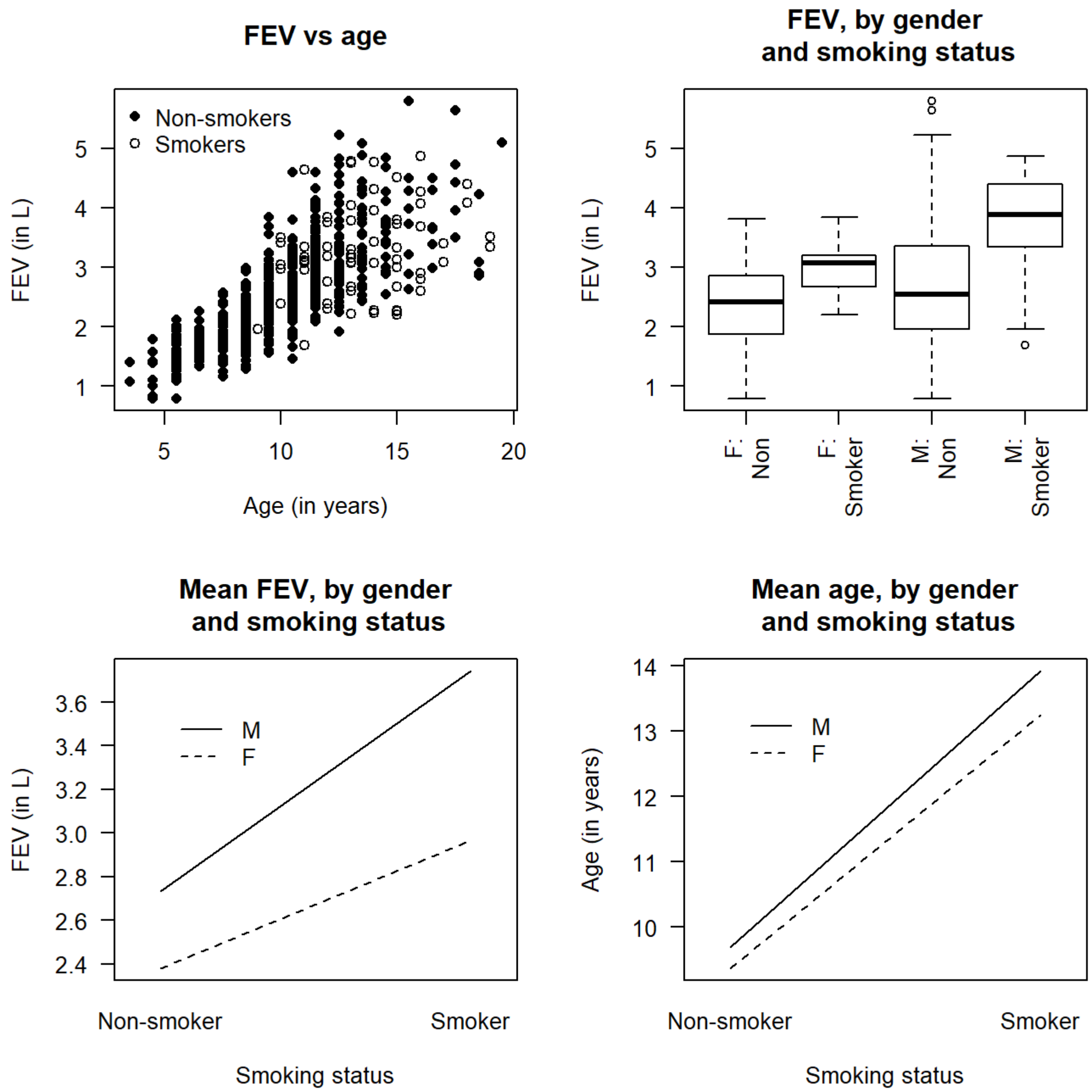
```
##
## Non-smoker     Smoker
##        589         65
```

1.3 Plotting data

Univariate



Multivariate



To make any further progress quantifying the relationship between the variables, mathematics is necessary to create a *statistical model*.

1.4 Factors

Concept of contrasts, “dunny vars”

```
contrasts(lungcap$Gender)
```

```
##   M
## F 0
## M 1
```

the contrasts reference level is arbitrary (alphabetical) unless you set it

```
contrasts( relevel( lungcap$Gender, "M") ) # Now, M is the ref. level
```

```
##   F
## M 0
## F 1
```

```
lungcap$Smoke <- factor(lungcap$Smoke,
  levels=c(0, 1),
  labels=c("Non-smoker","Smoker"))

contrasts(lungcap$Smoke)
```

```
##            Smoker
## Non-smoker      0
## Smoker          1
```

1.5 Statistical models

random component Describes the distribution of the dependent variable

systematic component Describes a “statistical model”

$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$
Here:
μ_i is the predicted value of an observation (y_i)
β are regression estimates, intercept and slopes
x are the parameter values for Age, Ht, Gender, and Smoke

Assumptions
$var[y_i] = \sigma^2$ **constant variance** (likely true given these data?)
$y_i \sim N(\mu_i, \sigma^2)$ **Gaussian residuals** (merely popular…)

1.6 Regression models

- linear regression (e.g. assumes constant variance, Gaussian, etc.)
- generalized linear model (GLM: other options for modelling variation and dist.)

NB linear regression *is* GLM, in the specific case where we model constant var + Gaussian residuals

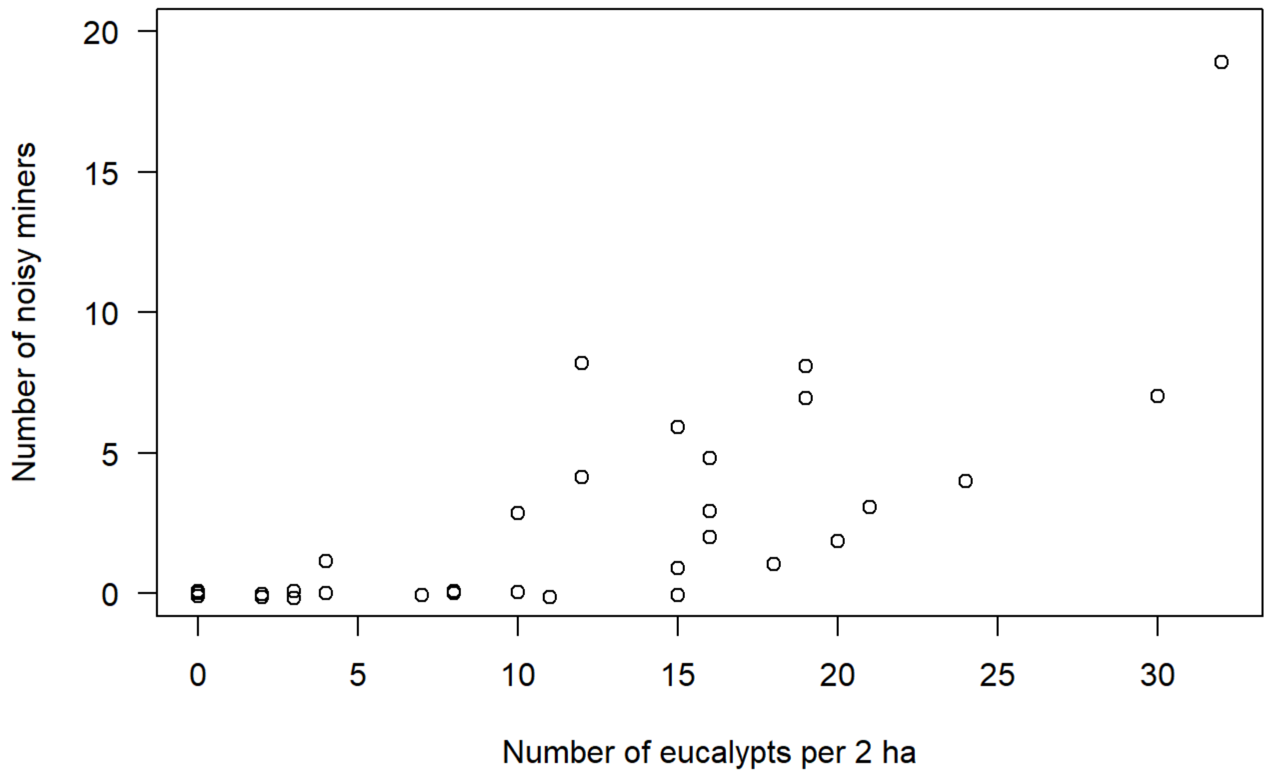
---

Poisson model for count data

```
data(nminer)
head(nminer)
```

```
##   Miners Eucs Area Grazed Shrubs Bulokes Timber Minerab
## 1      0    2   22      0      1     120     16       0
## 2      0   10   11      0      1      67     25       0
## 3      1   16   51      0      1      85     13       3
## 4      1   20   22      0      1      45     12       2
## 5      1   19    4      0      1     160     14       8
## 6      1   18   61      0      1      75      6       1
```

```
plot( jitter(Minerab) ~ Eucs, data=nminer, las=1, ylim=c(0, 20),
 xlab="Number of eucalypts per 2 ha", ylab="Number of noisy miners" )
```



1.8 All Models Are Wrong, but Some Are Useful

Good quote
* Prediction versus understanding
* Complexity, *Occam's Razor*
* Experiments versus observational study

Problems

1.1. The plots in Fig. 1.7 (data set: paper) show the strength of Kraft paper [7, 8] for different percentages of hardwood concentrations. Which systematic component, if any, appears most suitable for modelling the data? Explain.

Cubic - better than quadratic, simpler than Quartic

1.2. The plots in Fig. 1.8 (data set: heatcap) show the heat capacity of solid hydrogen bromide y measured as a function of temperature x [6, 16]. Which systematic component, if any, appears best for modelling the data? Explain.

Cubic - better than linear and simpler than Quadratis or Quartic

1.3. Consider the data plotted in Fig. 1.9. In the panels, quadratic, cubic and quartic systematic components are shown with the data. Which systematic component appears best for modelling the data? Explain.
Cubic
The data are actually randomly generated using the systematic component $\mu = 1+10\exp(-x/2)$ (with added randomness), which is not a polynomial at all. Explain what this demonstrates about fitting systematic components.
Demonstrates trial and error in explaining model variation!

1.4. Consider the data plotted in Fig. 1.10 (data set: toxo). The data show the proportion of the population y testing positive to toxoplasmosis against the annual rainfall x for 34 cities in El Salvador [5]. Analysis suggests a cubic model fits the data reasonably well (though substantial variation still exists). What important features of the data are evident from the plot? Which of the plotted systematic components appears better? Explain.
Uneven number of observations across range of rainfall. More extreme cubic linear regression seems overfitted on low data density compared to "gentler" fit of glm. Also range/extrapolation issue.

1.5. For the following systematic components used in a regression model, determine if they are appropriate for regression models linear in the parameters, linear regression models, and/or generalized linear models. In all cases, β_j refers to model parameters, μ is the expected value of the response variable, while x, x_1 and x_2 refer to explanatory variables.
Probably 2,3,4 okay. 1 definitely non-linear!

1.6

1. Use names() to determine the names of the variables in the data frame.

```
library(GLMsData)
data(turbines)
names(turbines)
```

```
## [1] "Hours"    "Turbines" "Fissures"
```

---

2. Determine which variables are quantitative and which are qualitative.

```
str(turbines)
```

```
## 'data.frame':    11 obs. of  3 variables:
##  $ Hours   : int  400 1000 1400 1800 2200 2600 3000 3400 3800 4200 ...
##  $ Turbines: int  39 53 33 73 30 39 42 13 34 40 ...
##  $ Fissures: int  0 4 2 7 5 9 9 6 22 21 ...
```

None are qualitative, but possibly Hours should be - discuss?

---

3. For any qualitative variables, define appropriate dummy variables using treatment coding.
   ...

---

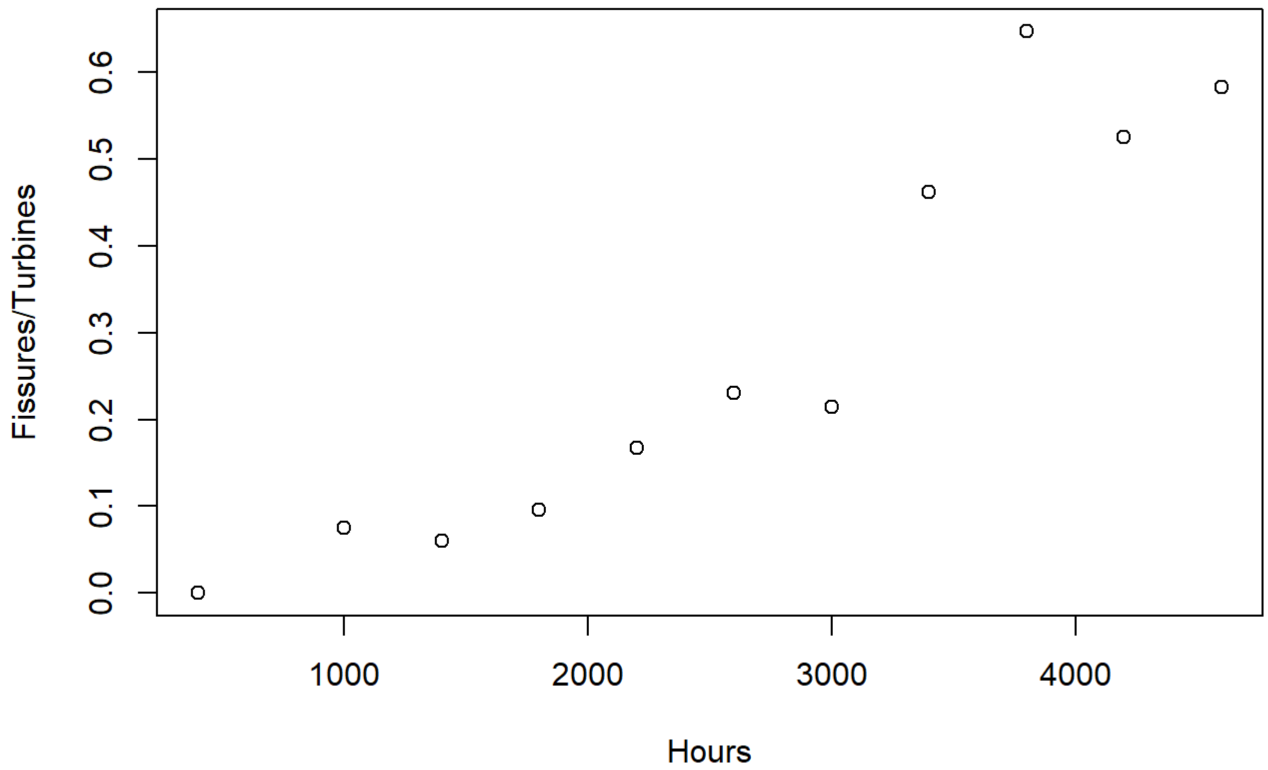
4. Use r to summarize each variable.

```
summary(turbines)
```

```
##      Hours         Turbines        Fissures
##  Min.   : 400   Min.   :13.00   Min.   : 0.000
##  1st Qu.:1600   1st Qu.:33.50   1st Qu.: 4.500
##  Median :2600   Median :39.00   Median : 7.000
##  Mean   :2582   Mean   :39.27   Mean   : 9.636
##  3rd Qu.:3600   3rd Qu.:41.00   3rd Qu.:15.000
##  Max.   :4600   Max.   :73.00   Max.   :22.000
```

5. Use r to create a plot of the proportion of failures (turbines with fissures) against run-time.

```
plot(Fissures/Turbines ~ Hours, data = turbines)
```



6. Determine the important features of the data evident from the plot.
   Not quite linear. Increasing variance with Hours. Described as an experiment, number hours running manipulated?

7. Would a linear regression model seem appropriate for modelling the data? Explain.
   Probably not inappropriate… violates some assumptions though

8. Read the help for the data frame (use ?turbines after loading the GLMsData package in r), and determine whether the data come from an observational or experimental study, then discuss the implications.

This is an experiment. Therefore we can make prediction about fissures as a function of hours run, and infer causation

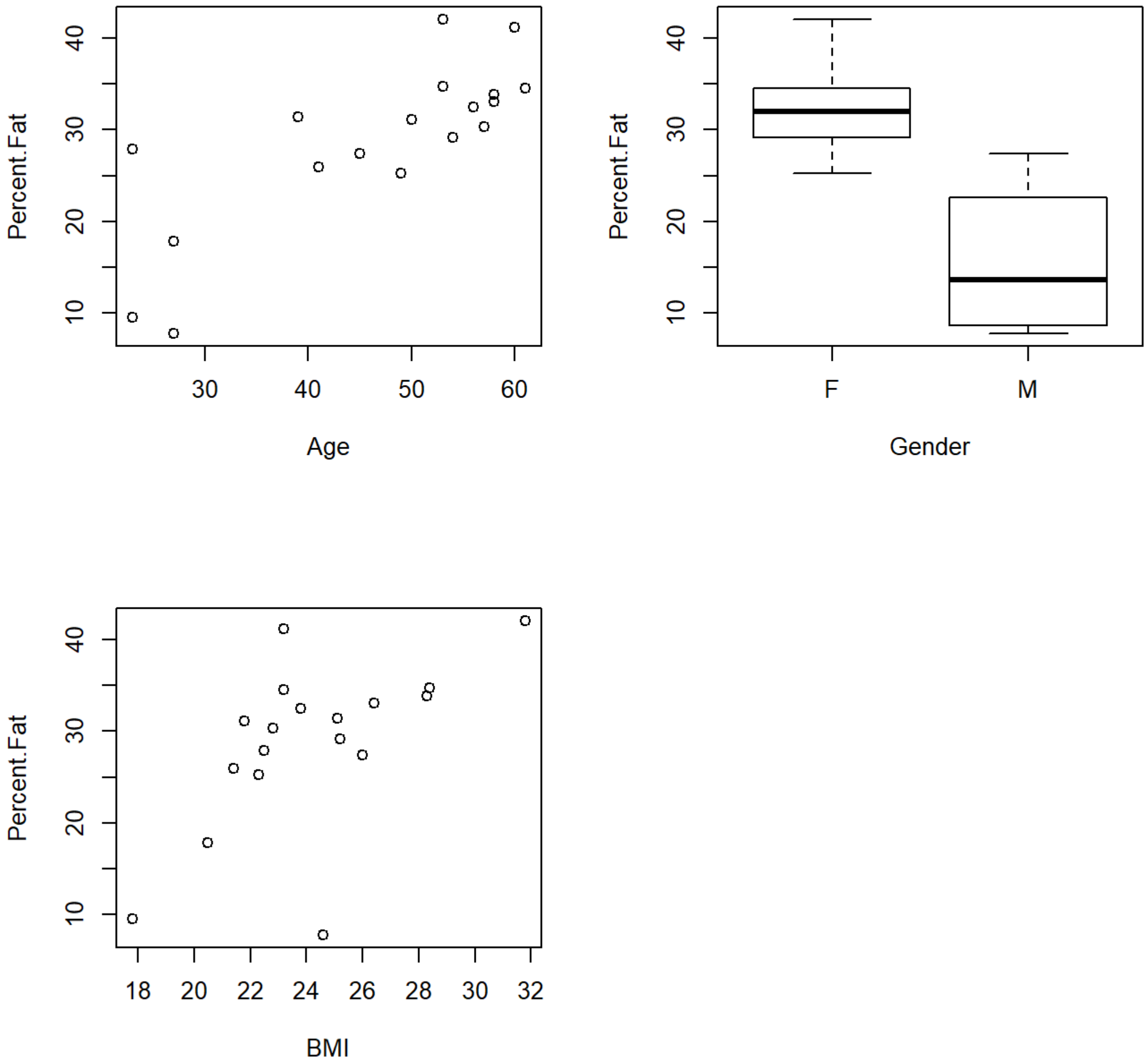
---

1.7

```
## [1] "Age"         "Percent.Fat" "Gender"      "BMI"
```

```
##   M
## F 0
## M 1
```

```
##       Age         Percent.Fat    Gender      BMI
##  Min.   :23.00   Min.   : 7.80   F:14   Min.   :17.80
##  1st Qu.:39.50   1st Qu.:26.27   M: 4   1st Qu.:22.35
##  Median :51.50   Median :30.70          Median :23.50
##  Mean   :46.33   Mean   :28.61          Mean   :24.17
##  3rd Qu.:56.75   3rd Qu.:33.60          3rd Qu.:25.80
##  Max.   :61.00   Max.   :42.00          Max.   :31.80
```


Percent.Fat
10
20
30
40
30
40
50
60
Age
Percent.Fat
10
20
30
40
F
M
Gender
Percent.Fat
10
20
30
40
18
20
22
24
26
28
30
32
BMI