Chapter 2

Name	Content
TYPE	notes
BOOK	An Introduction to Statistical Learning
AUTHORS	Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani
PUBLISHER	Springer

// Note: //
The .Rmd file was written and compiled using the Atom.io text editor. If you download the .Rmd file and try to open it in R Studio the LaTeX equations *will not* display properly. This has a easy fix: delete the spaces between the dollar signs ($).

Unfortunately, the R code chunks will not run properly in R Studio when downloading this .Rmd file. That is because the parameters listed inside the curly braces, {}, are incorrect. This fix is a little more time intensive, but is possible. For R studio the parameters take the form:

- {r loaddata, attr.source='.numberLines'}

For Atom (using the Hydrogen and markdown-preview-enhanced packages), the paramaters take the form:

- {r id="loaddata", .line-numbers}

a useful guide for using R in Atom can be found here: R in Atom

- how to use Atom with Rmarkdown: Rmarkdown in Atom

why? Atom has native Github integration, the interface is cleaner, and you're represented by an adorable octocat.You don't need to use Atom. In this repo I've also included the PDF version of these notes. :)
//

```
# the data from the book can be downloaded using install.packages("ISLR"). It's then loaded using the line below.
library(MASS)
```

2.1 What is Statistical Learning?

X denotes the input variable (aka: predictor or independent variable) Y denotes the output variable (aka: response or dependent variable)

The relationship between X and Y can be written as: $Y = f(X) + \mathcal{E}$

f is a fixed, but unknown function of X and $\mathcal{E}$ is the error term. $\mathcal{E}$ is independent of X and has a mean of 0. The function f may take more than one input variable. (e.g. income (y) as function of education (X_1) and seniority (X_2)).

Statistics is all about ways to estimate f

2.1.1 Why Estimate f?

1. Prediction: this is when we know the values of X, but can't easily determine Y.

- $\hat{y} = \hat{f}\,(x)$
- $\hat{Y}$ is the resulting predicting for Y (aka the predicted response)
- $\hat{f}$ is the estimate for f. It is a *black box* - where we don't really care about $\hat{f}$ as long as it gives accurate predictions for Y.

Generally, $\hat{f}$ will not be a perfect estimate f, as a result the inaccuracy will introduce some error.

reducible error: can be reduced to improve the accuracy of $\hat{f}$ by using better statistical methods. *irreducible error:* since Y is a function of $\mathcal{E}$, not all of the error can be reduced. Therefore there will always be $\mathcal{E}$

$\mathcal{E}$ is not zero because it might include variables that are useful in predicting Y and since we don't measure these unincluded variables they can't be predicted using f. $\mathcal{E}$ may also contain unmeasurable variation, which also can't be predicted using f

So if we have estimate $\hat{f}$ and predictors X we get the prediction: $\hat{Y}$ = $\hat{f}\,(X)$. If we assume that $\hat{f}$ and X are fixed:

$$E(Y - \hat{Y})^2 = E[f(X) - \mathcal{E} - \hat{f}\,(X)]^2 = [f(X) - \hat{f}\,(X)]^2 + var(\mathcal{E})$$

$E = (Y - Y^2)$ is the expected value of the squared difference between the predicted value and actual value of Y. $Var(\mathcal{E})$ is the variance associated with the error term $\mathcal{E}$

The irreducible error gives an upper bound on the accuracy of our prediction for Y & will almost always be unknown in practice.

2. Inference *Inference* is when we want to know how Y is affected by change in the predictors, $X_1 \dots X_p$, but aren't necessarily interested in making predictions for Y.

- the goal is to understand the relationship between X and Y. How Y changes as a function of $X_1 \dots X_p$
- $\hat{f}$ can't be treated as a *black box* because we have to know its exact form.
- linear models are useful for inference.

Inference is useful for:

- determining which predictors are associated with the outcome.
- determining the relationship between the outcome and each predictor.
- determining whether the relationship between Y and each predictor can be summarized using a linear equation or whether the relationship between the two is more complicated.

2.1.2 How do we Estimate f?

n is the number of data points or observations we have. *training data* is a subset of the data we have that we use to train (or teach) the method how to estimate f. We apply a statistical learning method to the training data in order to estimate f.

x_{ij} is the value of the jth predictor for observation i. i = 1, 2,…, n and j = 1, 2,…,p y_i is the response variable for ith observation. The training data would be $(x_i, y_i), (x_2, y_2), \dots (x_n, y_n)$, where $x = (x_{i1}, xi2, \dots, x_{ip})^T$.

The goal is to find $\hat{f}$ such that $Y \approx \hat{f}\,(X)$ for any observation (X, Y)

There are two statistical learning methods we can use: 1. *Parametric:* to estimate f we only need to estimate one set of parameters.

- the problem is that it will usually not match the true unknown form of f
- if the model is too far off from the true f (or the f using all the observations), the estimate will be poor
- to solve poor fit, we can use more flexible models. But more flexible models requires estimating more parameters.
- more complex modesl can lead to *overfitting:* which means they follow the errors too closely.
- these involve a two-step model-based approach.

Step 1 We make an assumption of f's form. For example, if f is linear: $f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ If f is linear, you only need to estimate the coefficients $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_p X_p$

Step 2 We use the training data to *fit* or *train* the model. For the linear model we want to estimate:
$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

One method of fitting the model is *(ordinary) least squares*

2. *Non-parametric:*

- do not make explicit assumptions about the functional form of f.
- these methods try to get as close to the data points without being too rough.
- since they don't assume particular form of f, they can can fit a wider range of shapes for f.
- will fit the data better since it does not assume the form of f.
- requires substantially more obeservations in order to get an accurate estimate for f than parametric approaches.

2.1.3 The Trade-off Between Prediction Accuracy and Model Interpretability

Some methods are less flexible because they can produce only a small range of shapes to estimate f (*e.g.* Linear regression can only create linear functions.)

- less flexible models are better for inference because they are more interpretable.
- it's easier to understand the relationship between Y and $X_1, X_2, \ldots X_p$ More flexible models include *thin plate splines* can generate a wider range of possible shapes to estimate f.

2.1.4 Supervised vs. Unsupervised Learning

Supervised Learning for each observation of the predictor measurements $x_i, i = 1, \ldots, n$ there is an associated response to the measurement y_i.

Unsupervised Learning is more complicated because for every observation $i = 1, \ldots, n$ there's a vector of measurements x_i, but no associated response y_i.

- it is called unsupervised because we have no response variable y to supervise our analysis.