ISLR Notes

TBD

2021

Contents

About

Notes and solutions for the exercises in the book: *An Introduction to Statistical Learning with Applications in R (1st edition)* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (website: https://www.statlearning.com/)

License

Chapter 1

Introduction

1.1 An Overview of Statistical Learning

> "Statistical learning refers to a vast set of tools for understanding data."

- Supervised: Using statistical models to **predict** or **estimate outputs** based on **inputs**.
- Unsupervised: Finding relationships between variables and structure in the data

1.2 Data sets

Example data used in the book

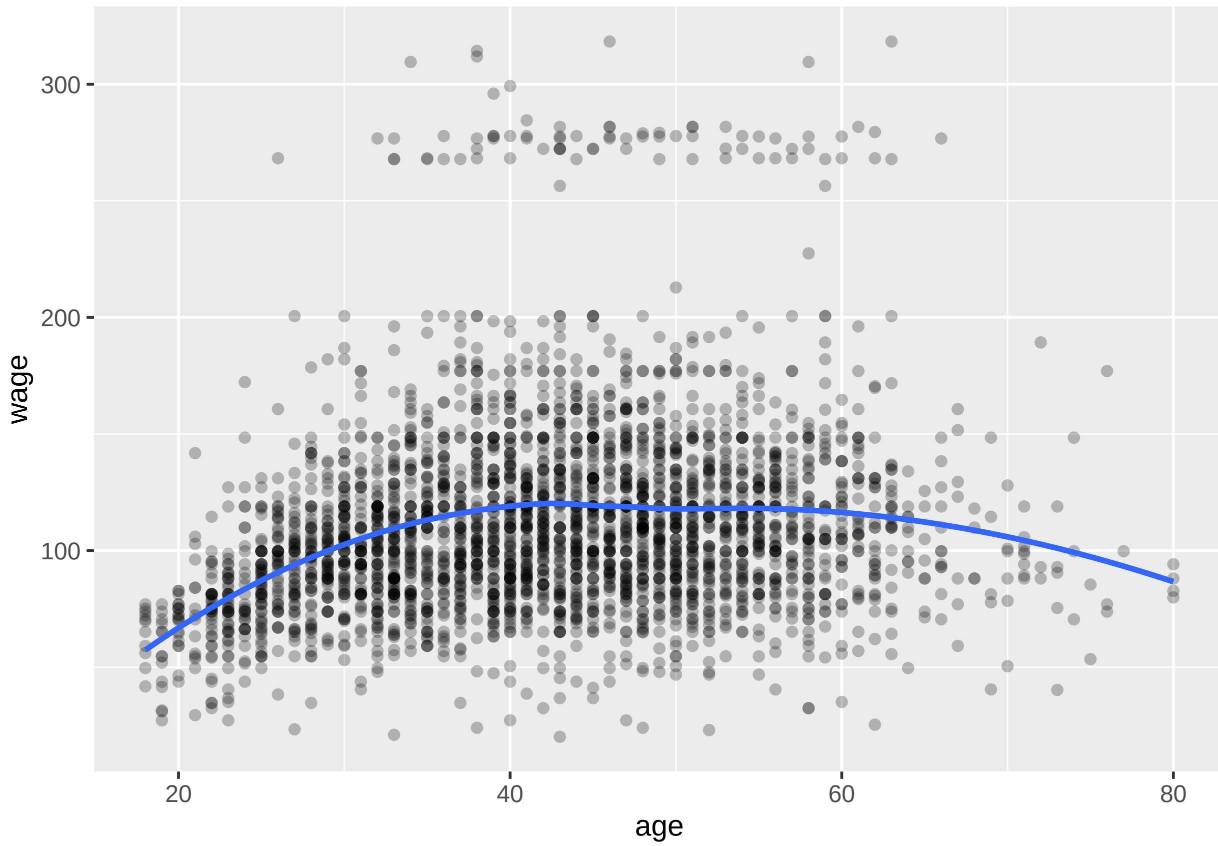
- Wages
- Stock Market Data
- Gene Expression Data

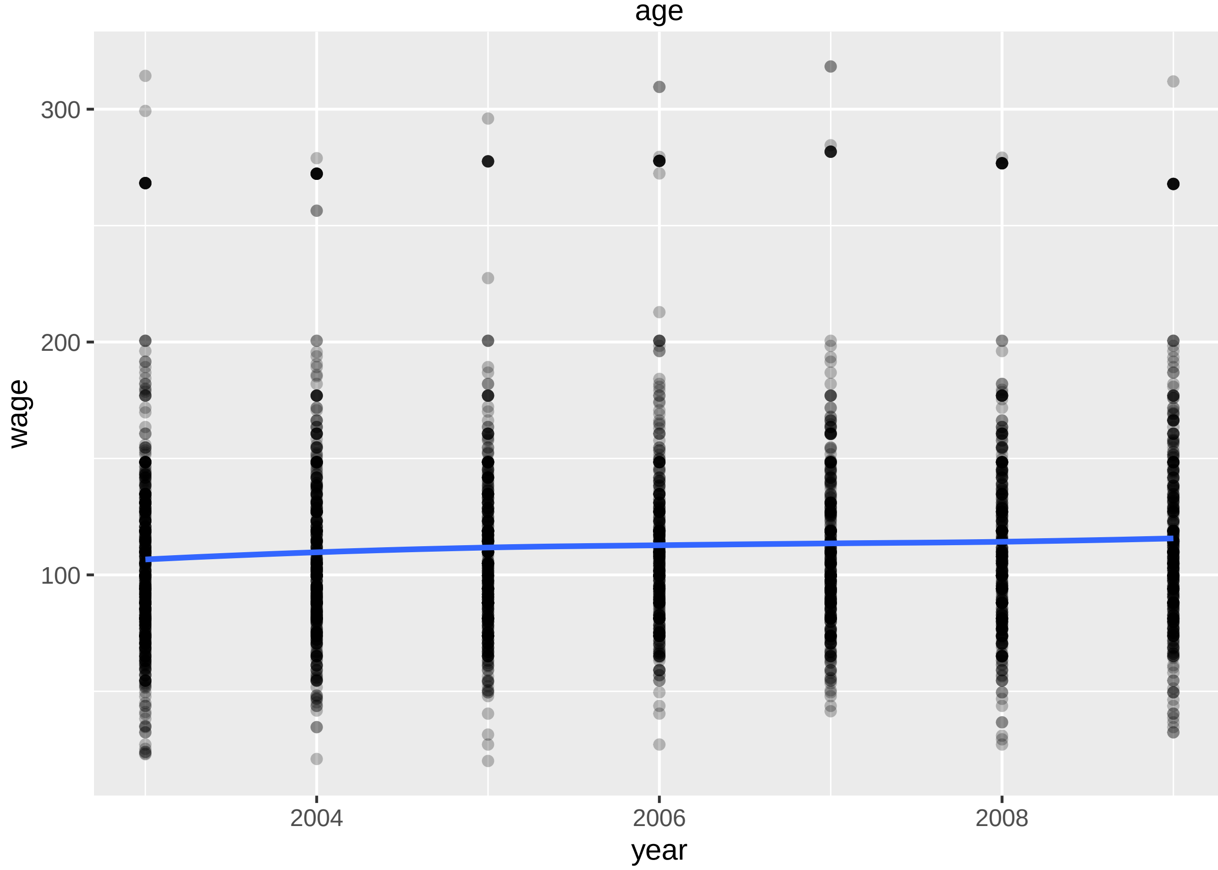
1.2.1 Wages

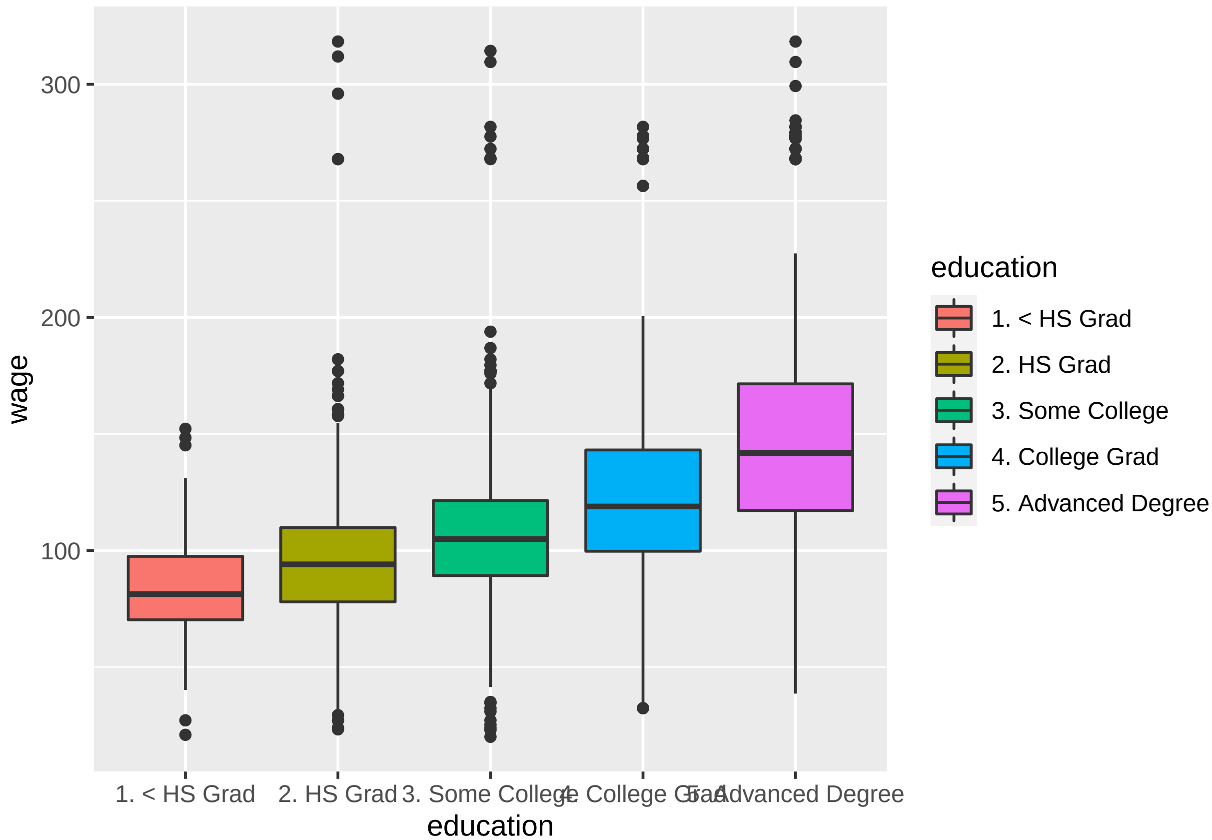
Used for regression problem examples such as predicting wage based on age and education

```
glimpse(Wage)
```

```
## Rows: 3,000
## Columns: 11
## $ year       <int> 2006, 2004, 2003, 2003, 2005, 2008, 2009, 2008, 2006, 2004,~
## $ age        <int> 18, 24, 45, 43, 50, 54, 44, 30, 41, 52, 45, 34, 35, 39, 54,~
## $ maritl     <fct> 1. Never Married, 1. Never Married, 2. Married, 2. Married,~
## $ race       <fct> 1. White, 1. White, 1. White, 3. Asian, 1. White, 1. White,~
## $ education  <fct> 1. < HS Grad, 4. College Grad, 3. Some College, 4. College ~
## $ region     <fct> 2. Middle Atlantic, 2. Middle Atlantic, 2. Middle Atlantic,~
## $ jobclass   <fct> 1. Industrial, 2. Information, 1. Industrial, 2. Informatio~
## $ health     <fct> 1. <=Good, 2. >=Very Good, 1. <=Good, 2. >=Very Good, 1. <=~
## $ health_ins <fct> 2. No, 2. No, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Yes, 1. Ye~
## $ logwage    <dbl> 4.318063, 4.255273, 4.875061, 5.041393, 4.318063, 4.845098,~
## $ wage       <dbl> 75.04315, 70.47602, 130.98218, 154.68529, 75.04315, 127.115~
```


wage
300
200
100
20
40
60
80
age


wage
300
200
100
2004
2006
2008
year



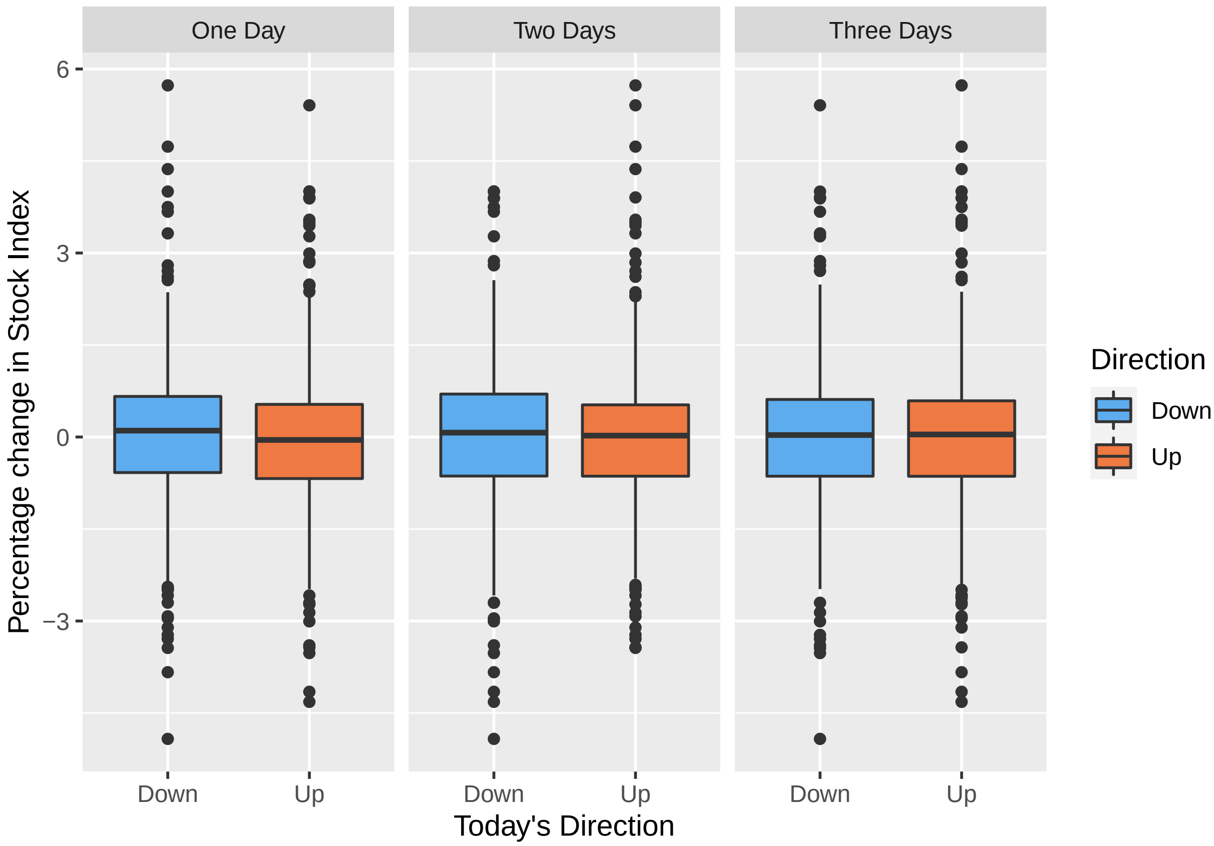
1.2.2 Stock Market Data

Used for classification problem examples with categorical or qualitative output, such as predicting whether a stock index will either increase or decrease on any given day.

Daily percentage change of S&P 500 stock index and 5 prior days

```
glimpse(Smarket)
```

```
## Rows: 1,250
## Columns: 9
## $ Year      <dbl> 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, 2001, ~
## $ Lag1      <dbl> 0.381, 0.959, 1.032, -0.623, 0.614, 0.213, 1.392, -0.403, 0.~
## $ Lag2      <dbl> -0.192, 0.381, 0.959, 1.032, -0.623, 0.614, 0.213, 1.392, -0~
## $ Lag3      <dbl> -2.624, -0.192, 0.381, 0.959, 1.032, -0.623, 0.614, 0.213, 1~
## $ Lag4      <dbl> -1.055, -2.624, -0.192, 0.381, 0.959, 1.032, -0.623, 0.614, ~
## $ Lag5      <dbl> 5.010, -1.055, -2.624, -0.192, 0.381, 0.959, 1.032, -0.623, ~
## $ Volume    <dbl> 1.1913, 1.2965, 1.4112, 1.2760, 1.2057, 1.3491, 1.4450, 1.40~
## $ Today     <dbl> 0.959, 1.032, -0.623, 0.614, 0.213, 1.392, -0.403, 0.027, 1.~
## $ Direction <fct> Up, Up, Down, Up, Up, Up, Down, Up, Up, Up, Down, Down, Up, ~
```



1.2.3 Gene Expression Data

Used for examples of clustering problems such as identifying related groups of cancer cells based on observed characteristics.

```
str(NCI60)
```

```
## List of 2
##  $ data: num [1:64, 1:6830] 0.3 0.68 0.94 0.28 0.485 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:64] "V1" "V2" "V3" "V4" ...
##   .. ..$ : chr [1:6830] "1" "2" "3" "4" ...
##  $ labs: chr [1:64] "CNS" "CNS" "CNS" "RENAL" ...
```

1.3 History

A brief timeline for the development of statistical learning

- 1800's *Linear Regression* (*Method of Least Squares*)
- 1936 *Linear Discriminant Analysis* developed to predict qualitative values
- 1940s *Logistic Regression* developed to predict qualitative values
- 1970s *Generalized Linear Models* including both logistic and linear regression
- 1980s *Classification and Regression Trees*
- 1986 *Generalized Additive Models*
- Present day (2001) *Machine Learning*

1.4 Other Considerations

"How Eugenics Shaped Statistics: Exposing the damned lies of three science pioneers.

1.5 Matrix Notation

Conventions used in the book

- n number of observations in a sample
- p number of variables
- **X** an $n \times p$ matrix
  - where x_{ij} represents the element in the ith row and the jth column.
  - x_i represents a single observation (row) as a vector with length p. Note that vectors are written vertically by convention in math notation.
  - $\mathbf{x}_j$ represents a single variable (column) as a vector with length n. Note that the bold face font is used to distinguish columns ($\mathbf{x}_3$) from rows (x_3).
- The ^{T} superscript operator denotes the transpose of a matrix or vector, where row and column indices are reversed such that the resulting matrix or vector will have p rows and/or n columns.

Examples

- A matrix of elements

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

- A row vector

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

- A column vector

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

- A matrix represented as a collection of column vectors

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j)$$

- A transposed matrix. Rows become columns and columns become rows

$$\mathbf{X}^T = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \dots & x_{pn} \end{pmatrix}$$

- A transposed row vector. Again, vector elements are listed vertically by default, so this presentation shows the new orientation.

$$x_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$$

- A matrix represented as a collection of row vectors

$$\mathbf{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}$$

Chapter 2

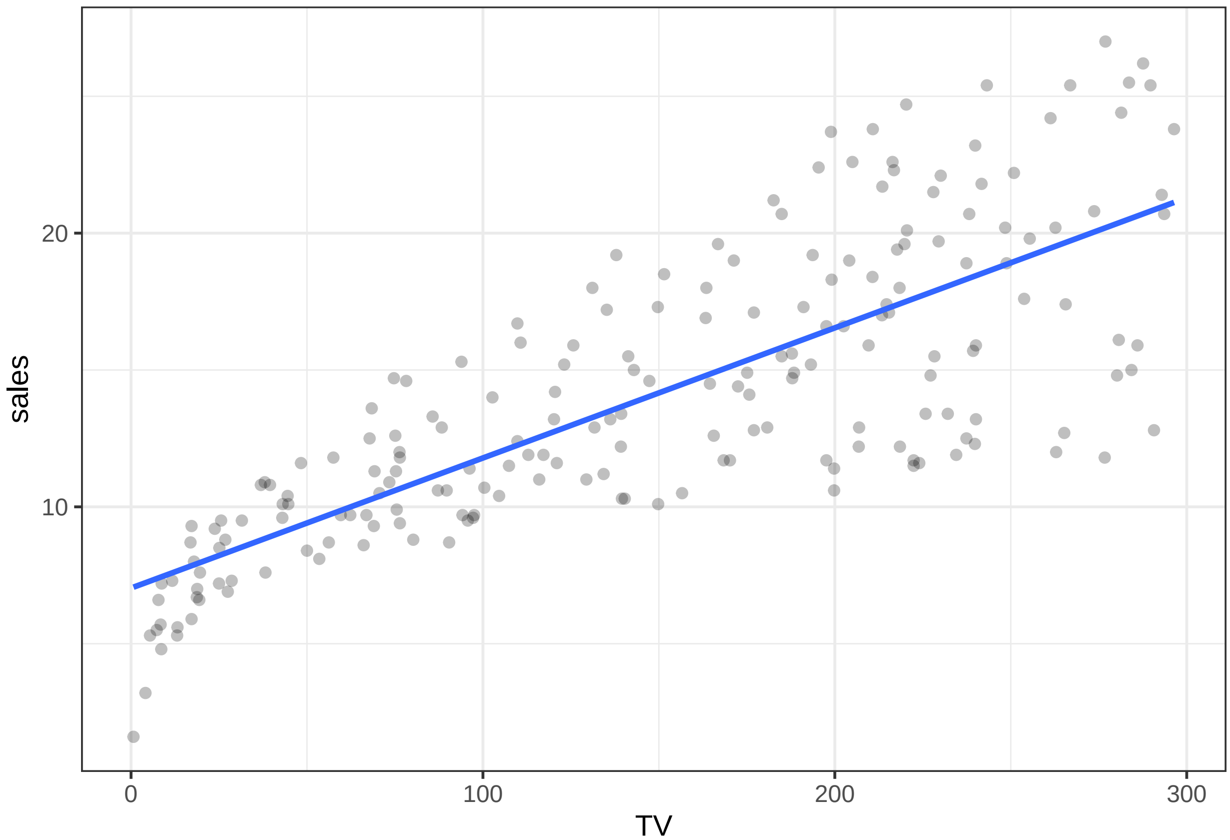
Statistical Learning

2.1 2.1 What Is Statistical Learning?

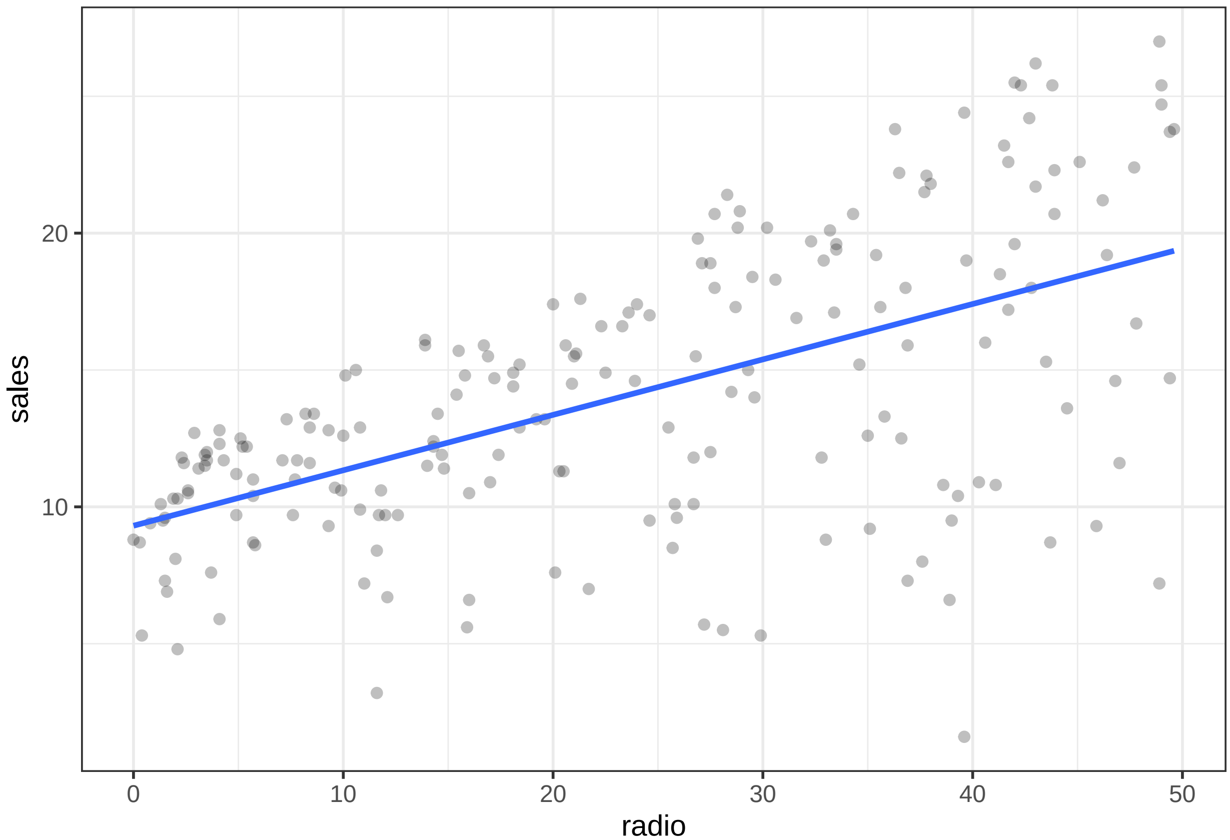
Motivating example: > Suppose that we are statistical consultants hired by a client to provide advice on how to improve sales of a particular product. ... our goal is to develop an accurate model that can be used to predict sales on the basis of the three media budgets.

```
## Rows: 200
## Columns: 4
## $ TV        <dbl> 230.1, 44.5, 17.2, 151.5, 180.8, 8.7, 57.5, 120.2, 8.6, 199.~
## $ radio     <dbl> 37.8, 39.3, 45.9, 41.3, 10.8, 48.9, 32.8, 19.6, 2.1, 2.6, 5.~
## $ newspaper <dbl> 69.2, 45.1, 69.3, 58.5, 58.4, 75.0, 23.5, 11.6, 1.0, 21.2, 2~
## $ sales     <dbl> 22.1, 10.4, 9.3, 18.5, 12.9, 7.2, 11.8, 13.2, 4.8, 10.6, 8.6~
```

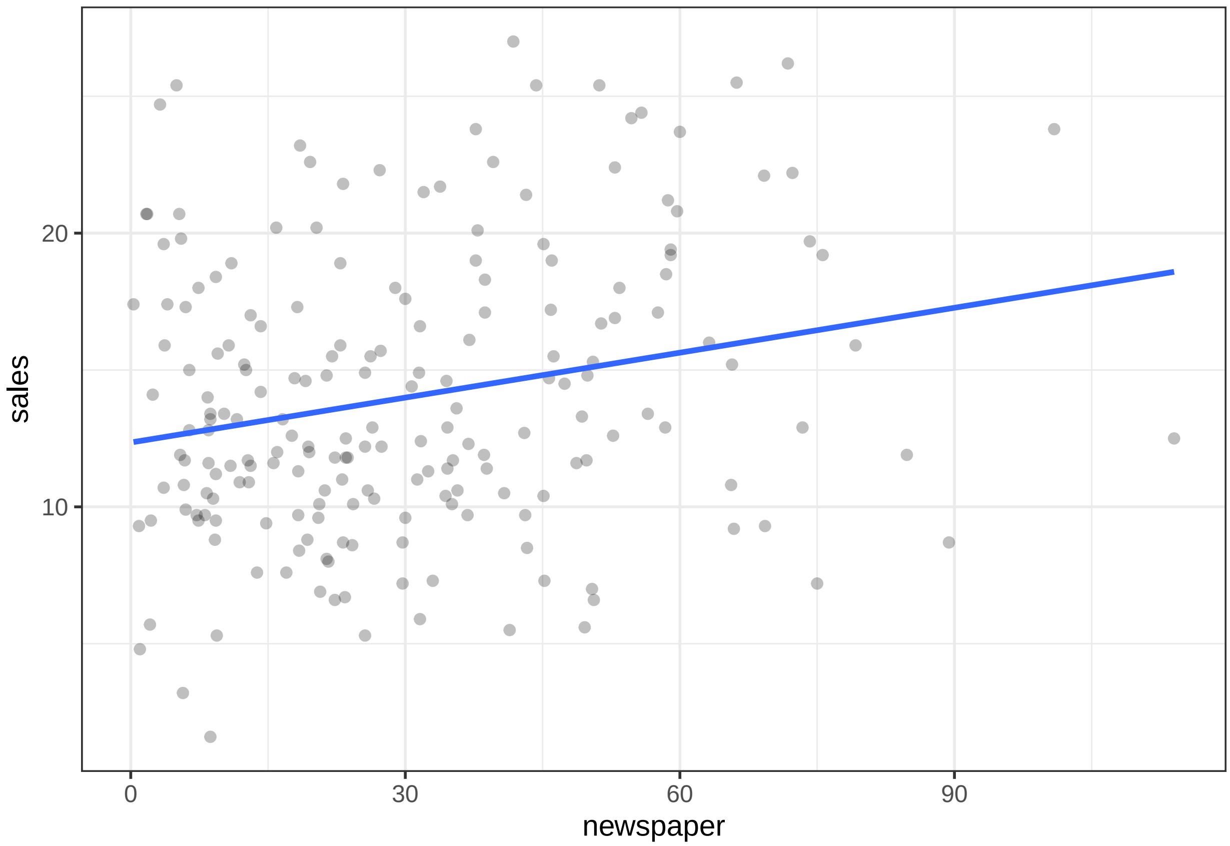
```
## Warning: Ignoring unknown parameters: point
```




```
## Warning: Ignoring unknown parameters: point
```




```
## Warning: Ignoring unknown parameters: point
```



Input Variables: These are the variables we knoW and can use to build our model. Also known as *predictors*, *independent variables*, or *features*. Denoted using the symbol X_n.

Output Variable: This is the variable we are trying to predict with the model. Also known as a *response*, or *dependent variable*. Typically denoted as Y.

More generally: $Y = f(X) + \epsilon$

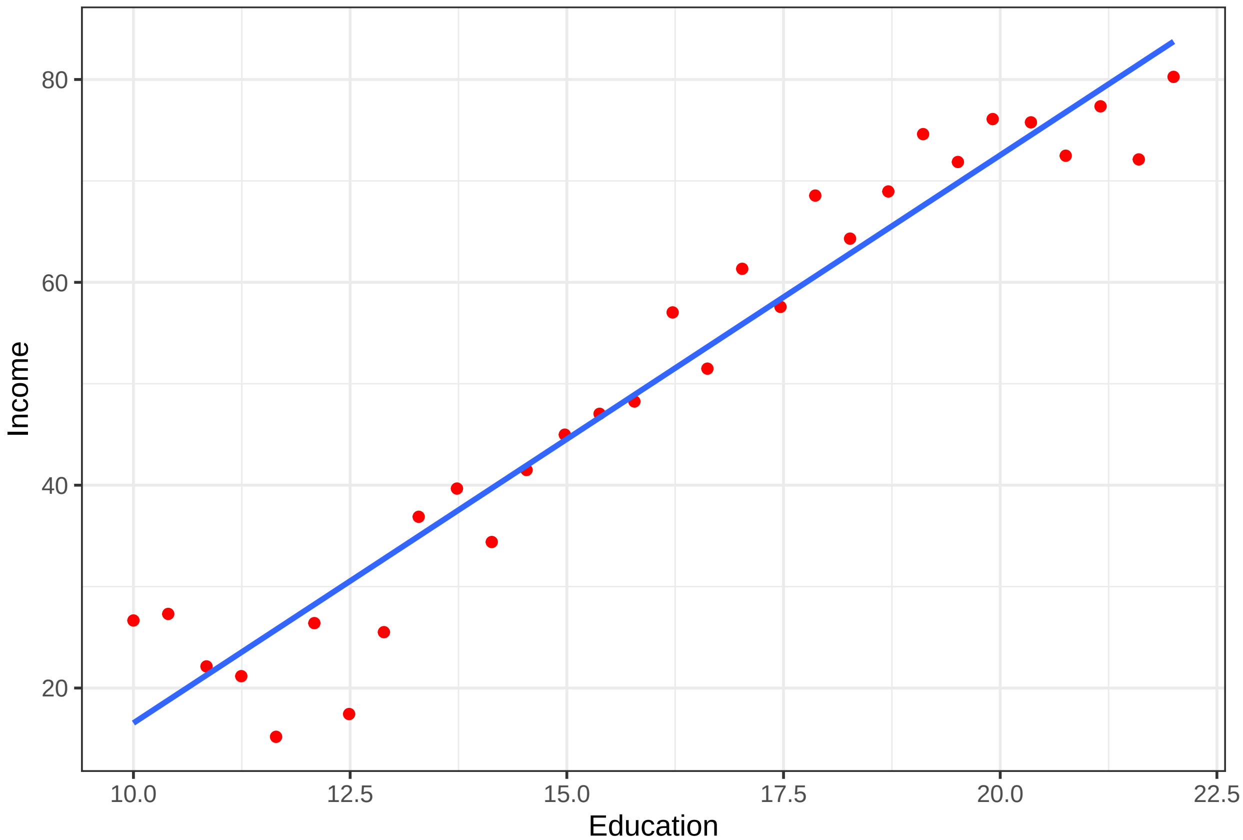
Where Y is the quantitative response and f is a function of $X_1, ..., X_p$ (of p different predictors) and ϵ is some random **error term**.

Assumptions:

- f is **systematic** in its relationship to Y
- ϵ is independent of X
- ϵ has mean zero

Another example: Income and education may appear related, but the exact relationship is unknown. Note that some of the observations are above the linear interpolated line, while some are below it. The difference is ϵ

```
## Warning: Ignoring unknown parameters: point
```



2.2 2.1.1 Why Estimate f?

There are two main reasons to estimate f:
- Prediction - Inference

2.2.1 Prediction

Consider: $\hat{Y} = \hat{f}(X)$

If X is known, we can predict $\hat{Y}$ by this equation. Don't be too concerned with the exact functional form of $\hat{f}$.

2.2.1.1 Terms:

- **reducible error**: This is error that comes with the model. We can address this error by improving the accuracy of the model.
- **irreducicle error**: This is error introduced to the model, because ϵ, by definition, cannot be explained by X

2.2.1.2 Ineference:

-

2.3 2.1.2 How Do We Estimate f?

2.4 2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

2.5 2.1.4 Supervised Versus Unsupervised Learning

2.6 2.1.5 Regression Versus Classification Problems

2.7 2.2 Assessing Model Accuracy

2.8 2.2.1 Measuring the Quality of Fit

2.9 2.2.2 The Bias-Variance Trade-Off

2.10 2.2.3 The Classification Setting

2.11 2.3 Lab: Introduction to R

2.12 2.3.1 Basic Commands

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

2.13 2.3.2 Graphics

2.14 2.3.3 Indexing Data

2.15 2.3.4 Loading Data

2.16 2.3.5 Additional Graphical and Numerical Summaries

2.17 2.4 Exercises

2.18 Conceptual

1.
2.
3.
4.
5.
6.
7.

2.19 Applied

8.
9.

10.

Chapter 3

Linear Regression

Chapter 4

Classification

Chapter 5

Resampling Methods

Chapter 6

Model Selection and Regularization

Chapter 7

Moving Beyond Linearity

Chapter 8

Tree Based Methods

Chapter 9

Support Vector Machines

Chapter 10

Unsupervised Learning