MATH 377 & 378 NOTES

Philippos Dimitroglou

Table of Contents

Math 377 – Advanced Probability and Statistics

Chapter 1 – Descriptive Statistical Modeling Case Study

Data	Observations collected from the likes of field notes, surveys, and experiments
Research question	The principal question the researchers hope to answer
Summary statistic	A single number summarizing a large amount of data

Chapter 2 – Data Basics

Case	An individual instance of data; usually represented as rows
Variables	Characteristics of data; usually represented as columns
Data matrix, data frame, tibble	How data is organized in tidy data
Tidy data	Data where each row of a matrix corresponds to a unique case, and each column corresponds to a variable.
Numerical variable	A variable which can take a wide range of numerical values where it is sensible to add, subtract, or take averages with those values
Discrete variable	Numerical variables that can take on only specific numerical values
Continuous variable	Numerical variables that can take on any numerical value
Categorical variable	A variable that represents a category
Levels	The possible values of a categorical variable
Ordinal variable	A categorical variable that suggests a ranking of levels; e.g. education level between high school, college, or graduate school
Associated variables, dependent variables	Two variables that show some kind of connection
Negatively associated	When the presence of one variable in a pair of associated variables makes the presence of the other less likely
Positively associated	When the presence of one variable in a pair of associated variables makes the presence of the other more likely
Independent variables	Two variables which are not associated

Chapter 3 – Overview of Data Collection Principles

Population	The group a research question targets
Sample	A subset of a population
Anecdotal evidence	Data collected in a haphazard fashion
Bias	When a sample is skewed to be unrepresentative of the population
Simple random sample	Subset of a population selected randomly from the population
Systemic sample	One case is sampled after letting a fixed number of others pass by
Non-response	When there is no response
Representative	Describes a sample that reflects the entire population within the sample
Non-response bias	A bias that results from those who responded not being representative of the population
Convenience sample	A sample of individuals who are easily accessible
Explanatory variable	The suspected variable that causes a change in the response variable
Response variable	The suspected variable that responds to the explanatory variable
Observational study	Collecting data in a way that does not interfere with how the data arises
Cohort	A group of similar individuals followed for a long duration in a study
Experiment	What researchers conduct when they want to test the possibility of a causal connection; there is interference in how the data arises
Randomized experiment	An experiment in which participants are assigned randomly between the treatment and control groups
Placebo	A fake treatment

Chapter 4 – Studies

Confounding variable	A variable correlated with both the response and explanatory variables
Prospective study	Identifies individuals and collects information as events unfold
Retrospective study	Collect data after events have taken place
Simple random sampling	A sample taken by randomly drawn from the population
Stratified sampling	Population is divided into groups called **strata**; then simple random sampling occurs within those groups
Cluster sampling	Observations are grouped into clusters, then randomly, some of the samplings are selected
Multistage sampling	Observations are grouped into clusters, then take a random sample within each selected cluster
Experiment	What researchers conduct when they want to test the possibility of a causal connection; there is interference in how the data arises
Randomized experiment	An experiment in which participants are assigned randomly between the treatment and control groups
Principles of Experimental Design	
Controlling	Ensuring consistency across possibly confounding variables; ex. In an experiment about medication, the pill is taken with same amount of water
Randomization	Researchers randomize people into treatment or control groups
Replication	Collecting a sufficiently large sample
Blocking	Grouping individuals based on confounding variables and then running experiment within these groups; ex. For a COVID vaccine study, grouping people by age

Chapter 5 – Numerical Data

Scatterplot	Provides a case-by-case view of data for two numerical variables
Mean	Measures the center of the distribution of data $x = \frac{x_1+x_2+\cdots+x_n}{n}$ for n cases
Point estimate	What the sample mean provides of the population mean
Weighted mean	A mean divided into groups so that unequal cases are not treated the same
Histogram	Plots showing binned counts plotted as bars
Data density	What histograms provide a view of; density corresponds to how high the bars are
Tail	Where the data trails off
Right skewed	When the tail of the data is to the right
Left skewed	When the tail of the data is to the left
Symmetric	When the data has equal tails on either side
Mode	Represented as a prominent peak in the distribution; the variable value the most cases have
Unimodal, bimodal, multimodal	Corresponding to one, two, or three or more modes
Variance	The sum of all deviations averaged $s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{n-1}$
Deviation	The distance of the data from its mean $(x - \bar{x})$
Standard deviation	The square root of the variance
Box plot	Summarizes data using five statistics: mean, median, interquartile range, first quartile, and third quartile
Median	Splits the data in half

Interquartile range (IQR)	Total length of box in box plot
First quartile	Lower boundary of box in box plot; 25% of data falls below value
Third quartile	Top boundary of box in box plot; 75% of data falls below value
Whiskers	Extensions of box plot that capture data outside of box; reach is never allowed to be more than $1.5 \times IQR$
Outliers	Unusual distant observations
Robust estimates	The median and IQR; observations where extreme observations have little effect on their values
Transformation	Rescaling of data using a function

Chapter 6 – Categorical Data

Factor	When categorical data that is a number is treated as categorical data
Contingency table	A table that shows the number of times a particular combination of data occurred
Row totals	Total counts across each row
Column totals	Total counts across each column
Marginal counts	Row and column totals
Joint counts	Values in the table
Frequency table	A table showing the counts of a single variable
Relative frequency table	Contingency table where counts are replaced with percentages or proportions
Bar plot	Displays a single categorical variable
Column proportions	Counts divided by their column totals
Conditioning	Restricting to only show data of specific elements
Segmented bar plot	Graphical display of contingency table information by columns
Mosaic plot	Graphical display of contingency table information by areas
Pie chart	Graphical display of contingency table information by percentages of the area of circles
Side by side box plot	Box plots split up by groups
Density plot	Compares data density across groups; like a box plot but with a smooth line

Chapter 7 – Probability Modeling Case Study

Parameters	Values that determine probabilities
Permutation, combinations	Counting methods used in probability calculations
Multiplication rule	Probability of x occurring equals $\frac{a}{b}$ where a is the number of sets of the desired outcome and b is the number of total possible outcomes Ex. Possibility of at least two people amongst a group of 18 having the same birthday is $$1 - \frac{\frac{365!}{(365-18)!}}{365^{18}}$$ which is 1 minus the probability that everyone will have different birthdays
Complementary probability	The probability of x is 1 minus the probability of the opposite of x.

Chapter 8 – Probability Rules

Sample space	The set S of all the possible results of the experiment
Outcome	An element of the sample space
Event	A subset of outcomes
Subset	A is considered a subset of B if all the outcomes of A are also contained in B
Intersection	All of the outcomes contained in both A and B
Union	All of the outcomes contained in A or B
Complement	All of the outcomes not contained in A
Probability	A number assigned to an event of outcome that describes how likely it is to occur
Probability model	Assigns a probability to each element of the sample space; can be thought of as a function mapping outcomes to a real number in $[0,1]$
Mutually exclusive	When events have disjoint, or no outcomes in common
Exhaustive	The union of all events is the sample space
Independent	Probabilities of event y after x is not influenced by x.
Equally likely scenarios	Each outcome is equally likely; thus $\mathbb{P}(A) = \frac{\#\ of\ outcomes\ in\ A}{\#\ of\ outcomes\ in\ S}$

Probability Axioms

Let S be the sample space of a random experiment and let A be an event where $A \subset S$. We have that

1. $\mathbb{P}(A) \geq 0$ (probability must be positive)
2. $\mathbb{P}(S) = 1$ (probability of all outcomes must sum to 1)

Where $\mathbb{P}(X)$ is the probability of X.

Probability Properties

1. $\mathbb{P}(\emptyset) = 0$
2. $\mathbb{P}(A') = 1 - \mathbb{P}(A)$
3. If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.
4. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$, can be generalized to more than two events; the intersection is subtracted because outcomes in both events get counted twice in the first sum
5. The Law of Total Probability: Let $B_1, B_2, \ldots, B_n$ be mutually exclusive and exhaustive. Then we have
$$\mathbb{P}(A) = \mathbb{P}(A \cap B_1) + \mathbb{P}(A \cap B_2) + \cdots + \mathbb{P}(A \cap B_n)$$
We also have Bayes' Rule which states $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B')$ which states A can be portioned into two parts: everything in A and B and everything in A and not in B.
6. DeMorgan's Laws
$$\mathbb{P}((A \cup B)') = \mathbb{P}(A' \cap B')$$
$$\mathbb{P}((A \cap B)') = \mathbb{P}(A' \cup B')$$

Counting Rules

Recall $n!$ is number of ways you can choose n objects

1. **Multiplication Rule 1 (Multiplication): Order Matters, Sample with Replacement**

   Multiply all numbers of possible outcomes together

2. **Multiplication Rule 2 (Permutation): Order Matters, Sampling Without Replacement**

   The following is the number of ways to select k objects from a group of n size (order matters)
$${}_nP_k = \frac{n!}{(n-k)!}$$

3. **Multiplication Rule 3 (Combination): Order Does not Matter, Sampling Without Replacement**

   The following is the number of ways to select k objects from a group of n (order does not matter)
$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}$$

Chapter 9 – Conditional Probability

Conditional probability - $P(A\ given\ B) = \ P(A|B) = \frac{P(A \cap B)}{P(B)}$

A and B are considered independent if and only if $P(A|B) = P(A)$ or $P(A)\ \cdot P(B) = P(A \cap B)$

Two events are independent if knowing one happened does not affect the probability of the other.

$P(A^C|B) = 1 - P(A|B)$ given does not change

Bayes' Rule: Let $B_1, B_2, \ldots, B_n$ be mutually exclusive and exhaustive events and let $P(A) > 0$. Then $P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$

Ex. Suppose a doctor has developed a blood test for a certain rare disease.

Sensitivity	The probability of detecting the disease for those who actually have it; probability test finds what it wants
Specificity	The probability of correctly identifying "no disease" for those who do not have it; probability that what test found was what it wanted

Lie with probability by equating joint probability with conditional probability when in reality, it is conditional probability that matters

Chapter 10 – Random Variables

Random variable	A variable that maps the events in the sample space to the real number line; random variable X is a one-to-one function $X: S \to \mathbb{R}$ where S is the sample space
Support	The sample space of X represented as S_X; the set of all the values X can take
Discrete random variable	Describes a support where X is a countable list of numbers
Continuous random variable	Describes a support that is a continuous interval
Mixed random variable	Describes a support that is both discrete and continuous
Distribution functions	Functions that describe the behavior of random variables
Probability mass function (pmf)	Let X be a discrete random variable. The probability mass function (pmf) of X, given by $f_X(x)$, is a function that assigns probability to each possible outcome of X. $$f_X(x) = P(X = x)$$
Cumulative distribution function (cdf)	Let X be a discrete random variable. The cumulative distribution function of X given by $F_X(x)$ is a function that assigns each value of X the probability that X takes that value or lower $$F_X(x) = P(X \leq x) = \sum_{y \leq x} f_X(y)$$
Moments	Values that summarize random variables with single numbers
Expectation	Let $g(x)$ be some function of a discrete random variable X. The expected value of $g(X)$ is given by $$E\big(g(X)\big) = \sum_{x} g(x) \cdot f_X(x)$$
Mean	The average value of the random variable

	$\mu_X = E(X) = \sum_x x \cdot f_X(x)$
Variance	Measure of spread of a random variable $\sigma_X^2 = Var(X) = E[(X-\mu_X)^2]$ Or in the discrete case $E[(X-\mu_X)^2] = \sum_x (x-\mu_X)^2 f_X(x)$

By the axioms of probability:

1. For all $x \in \mathbb{R}$, $0 \leq f_X(x) \leq 1$.
2. $\sum_x f_X(x) = 1$ where the x in the index of the sum simply denotes that we are summing across the entire domain or support of X.

$$F_X(x) = P(X \leq x) = \sum_{y \leq x} f_X(y)$$

Lemma:

Let X be a discrete random variable, and let a and b be constants. Then

$$E(aX+b) = aE(X) + b$$

and

$$Var(aX+b) = a^2 Var(X).$$

Chapter 11 – Continuous Random Variables

Probability density function (pdf)	Let X be a continuous random variable. The probability density function of X, given by $f_X(x)$ is a function that describes the behavior of X.
Cumulative distribution function (cdf)	Same principle as for a discrete variable, $$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\ dt$$
Expectation	Let $g(X)$ be any function of X. The expectation of $g(X)$ is found by $$E\big(g(X)\big) = \int_{S_X} g(x) f_X(x)\, dx$$
Mean	Let X be a continuous random variable. The mean of X or μ_X is simply $E(X)$. Thus $$E(X) = \int_{S_X} x f_X(x)\, dx$$
Variance	Similar to discrete case, we have $$\sigma_X^2 = Var(X) = E[(X-\mu_X)^2] = \int_{S_X} (x-\mu_X)^2 f_X(x)\, dx$$

Probability Density Function Properties

1. $F_X(x) \geq 0$
2. $\int_{S_x} f_X(x)\, dx = 1$
3. $P(X \in A) = \int_{x \in A} f_X(x)\, dx$ or another way to write this $P(a \leq X \leq b) = \int_a^b f_X(x)\, dx$

Properties 2 and 3 imply the area underneath a PDF represents a probability.

From the lemma last chapter, we have

$Var(X) = E(X^2) - E(X)^2 = \int_{S_X} x^2 f_X(x)\, dx - \mu_X^2$

Chapter 12 – Named Discrete Distributions

Discrete uniform distribution

A discrete random variable has the discrete uniform distribution if probability is evenly allocated to each value in the sample space

Let X be a discrete random variable with the uniform distribution. If the sample space is consecutive integers, this distribution is denoted as $X \sim Unif(a, b)$ where a and b represent the minimum and maximum of the sample space. The pmf of X is given by:

$$f_X(x) = \begin{cases} \frac{1}{b-a+1}, & x \in \{a, a+1, \dots, b\} \\ 0, & otherwise \end{cases}$$

Mean

$$E(X) = \frac{a+b}{2}$$

Variance

$$Var(X) = \frac{(b-a+1)^2 - 1}{12}$$

Binomial distribution

Let the random variable X represent the number of successes out of n repeated independent trials of a binary process that has a constant probability of success X is said to follow the binomial distribution with parameters n and p where p is the probability of success in each trial. The pmf of X is given by

$$f_X(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x \in \{0,1, \dots, n\}$ and 0 otherwise, representing the number of successes

Mean

$$E(X) = np$$

Variance

$$Var(X) = np(1-p)$$

Poisson process

Considers random processes where events occur according to some rate over time, eg. Arrivals at a retail register; assumes a consistent rate of arrival and a memoryless arrival process (the time until the next arrival is independent of time since the last arrival)

Let X be the number of arrivals in a length of time T where arrivals occur according to a Poisson process with an average of λ in length of time T (arrival / time). Then X follows a Poisson distribution with parameter λ:

$$X \sim Poisson(\lambda)$$

The pdf of X is given by

$$f_X(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0,1,2, \ldots$$

One unique feature of the Poisson distribution is that

$$E(X) = Var(X) = \lambda$$

Poisson distribution (random occurrences, fixed interval)

The distribution followed by number arrivals that occur in a specified amount of time

Hypergeometric distribution

Consider an experiment where k objects are to be selected from a larger, but finite, group consisting of m successes and n failures. This is similar to the binomial process, except in this case, results are selected without replacement. If the random variable X represents the number of successes selected in our sample size of k, then X follows a hypergeometric distribution with parameters m, n, and k.

The pmf of X is given by

$$f_X = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}, x = 0,1, \ldots, m$$

Mean

$$E(X) = \frac{km}{m+n}$$

Variance

$$Var(X) = k\frac{mn}{m+n}\frac{m+n-k}{m+n-1}$$

R Commands (suffix can be binom, unif, norm, pois, exp, hyper, gamma, beta,…)

d____

Equivalent to the probability mass function; used to find $\boldsymbol{P(X = x)}$;

p____

Equivalent to the cumulative distribution function; used to find $P(X \leq x)$;

q____

Inverse of the cumulative distribution function and will return a percentile; used to find $P(X \leq x) \geq p$;

r_____

This function is used to randomly generate values from the binomial distribution; it takes three inputs n (the number of values to generate), returns a vector containing generated values; used for simulations

Chapter 13 – Named Continuous Distributions

Uniform distribution

The distribution a continuous random variable takes on if probability density is constant. This distribution is commonly denoted as $U(a, b)$.

Let X be a continuous random variable with the uniform distribution. This is denoted as $X \sim Unif(a, b)$. The pdf of X is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & otherwise \end{cases}$$

Mean

$$E(x) = \frac{a+b}{2}$$

Variance

$$Var(X) = \frac{(b-a)^2}{12}$$

Exponential distribution (random interval, fixed occurrences)

Let X be the number of arrivals in a time interval T where arrivals occur according to a Poisson process with an average of λ arrivals per unit time interval. Now let Y be the time until the next arrival. Then Y follows the exponential distribution with parameter λ which has units of arrivals per unit time; $Y \sim Expon(\lambda)$. The pdf of Y is given by

$$f_Y(y) = \lambda e^{-\lambda y}, y > 0.$$

The cdf of Y is given by

$$F_Y(y) = P(Y \leq y) = 1 - e^{-\lambda y}$$

Mean

$$E(Y) = \frac{1}{\lambda}$$

Variance

$$Var(Y) = \frac{1}{\lambda^2}$$

Memoryless property

Time until next arrival is independent of time since last arrival; mathematically $P(X \geq s + t \,|X \geq s) = P(X \geq t)$

Gamma distribution

Generalization of the exponential distribution used to model wait times but without memoryless parameters. Suppose X is a random variable with the gamma distribution with shape parameter α and rate parameter λ; $X \sim Gamma(\alpha, \lambda)$. The pdf of X is given by

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{a-1} e^{\lambda x}, x > 0.$$

Mean	$E(X) = \frac{\alpha}{\lambda}$
Variance	$Var(X) = \frac{\alpha}{\lambda^2}$
Gamma function	$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ Notably $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ and if α is a non-negative integer, $\Gamma(\alpha) = (\alpha - 1)!$
Weibull distribution	Generalization of the exponential distribution used to model wait times. Suppose X is a random variable with the Weibull distribution with shape parameter α and scale parameter β. The pdf of X is given by $f_X(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, x \geq 0$
Normal or Gaussian distribution	Common distribution found in natural processes; bell curve is indicative of underlying normal distribution. Suppose X is a random variable with a normal distribution with mean μ and standard deviation σ. The pdf of X is given by $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$
Standard normal distribution	The distribution X follows when a random variable X is normally distributed with $\mu = 0$ and $\sigma = 1$. The pdf is denoted by $\phi(x)$. Any normally distributed random variable can be transformed to have the standard normal distribution. Let $X \sim Norm(\mu, \sigma)$. Then $Z = \frac{X - \mu}{\sigma} \sim Norm(0,1)$
Beta distribution	The domain of a random variable with the beta distribution is $[0,1]$. It is thus typically used to model proportions. The beta distribution has two parameters, α and β which are denoted as shape 1 and shape 2 respectively. Let $X \sim Beta(\alpha, \beta)$. The pdf of X is given by $f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}, 0 \leq x \leq 1$
Mean	$E(x) = \frac{\alpha}{\alpha + \beta}$
Variance	$Var(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$

Chapter 14 – Multivariable Distributions

Joint probability mass function

In the bivariate (depending on two variables) case, suppose X and Y are discrete random variables. The joint pmf is given by $f_{X,Y}(x,y)$ and represents

$$P(X=x,Y=y)=P(X=x\cap Y=y)$$

Marginal pmf

Gives pmf of individual variables according to

$$f_X(x)=\sum_{y\in S_Y} f_{X,Y}(x,y)$$

for the marginal pmf of X and

$$f_Y(y)=\sum_{x\in S_X} f_{X,Y}(x,y)$$

for the marginal pmf of Y.

Conditional pmf

Describes a discrete random variable given other random variables have taken particular values. In the bivariate case, the conditional pmf of X given $Y=y$, is denoted as $f_{X|Y=y}(x)$ and is found by

$$f_{X|Y=y}(x)=P(X=x|Y=y)=\frac{P(X=x,Y=y)}{P(Y=y)}=\frac{f_{X,Y}(x,y)}{f_Y(y)}$$

Joint probability density function

Describes joint probability of continuous random variables X and Y. Marginal pdfs are given by

$$f_X(x)=\int_{S_Y} f_{X,Y}(x,y)dy$$

for marginal pdf of X and

$$f_Y(y)=\int_{S_X} f_{X,Y}(x,y)dx$$

for marginal pdf of Y.

Conditional pdf

Found same way as conditional pdf in discrete case with notable difference for equality cases;

$$f_{X>x|Y=y}(x)=\int_{s_X}\frac{f_{XY}(x,y)}{f_Y(y)}dx$$

Where s_X is the subset of S_X that corresponds to the condition of X

Joint Probability Rules

1. $\sum_{x \in S_x} \sum_{y \in S_y} f_{X,Y}(x,y) = 1$ (discrete)
   $\int_{S_X} \int_{S_Y} f_{X,Y}(x,y)dydx = 1$ (continuous)
2. $0 \leq f_{X,Y}(x,y) \leq 1$

Chapter 15 – Multivariate Expectation

Expectation moment

Let X and Y be discrete random variables with joint pmf $f_{X,Y}(x,y)$. Let $g(X,Y)$ be some function of X and Y. Then

$$E[g(X,Y)] = \sum_x \sum_y g(x,y) f_{X,Y}(x,y)$$

In the case of continuous random variables with joint pdf $f_{X,Y}(x,y)$ the expectation becomes

$$E[g(X,Y)] = \int_x \int_y g(x,y) f_{X,Y}(x,y) dy dx$$

Mean (discrete)

Let X be a discrete random variable in a given joint pmf. Then

$$E(X) = \sum_x \sum_y x f_{X,Y}(x,y) = \sum_x x \sum_y f_{X,Y}(x,y) = \sum_x x f_X(x)$$

$$E(X^2) = \sum_x \sum_y x^2 f_{X,Y}(x,y) = \sum_x x^2 \, f_X(x)$$

$$E(XY) = \sum_x \sum_y xy f_{X,Y} \, (x,y)$$

Mean (continuous)

As with past definitions, this definition simply changes the summations in the discrete case into integrals over the sample space

Variance

$$Var(X) = E(X^2) - E(X)^2$$

Covariance

Let X and Y be two random variables. The covariance between X and Y is denoted as $Cov(X,Y)$ and is found by

$$Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$$

which simplifies into

$$Cov(X,Y) = E(XY) - E(X)E(Y)$$

Covariance can be positive or negative, with a positive covariance implying a direct linear relationship between variables and a negative covariance implying an inverse linear relationship. A zero covariance implies the random variables do not have a linear relationship, but not necessarily that they are independent.

Correlation

Let X and Y be two random variables. The correlation between X and Y is found from

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Covariance depends on the scales of the random variables involved. Thus, comparing covariance across sets variables is difficult. Correlation scales covariance so it can be compared.

Variance of sums

Let X and Y be two random variables. Then

$$Var(X+Y) = E\left[\left(X+Y-E(X+Y)\right)^2\right] = E[(X+Y)^2] - [E(X+y)]^2$$

which simplifies to

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

Independence

Two random variables X and Y are said to be independent if their joint pmf/pdf is the product of their marginal pmfs/pdfs according to

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

If X and Y are independent, then $Cov(X,Y) = 0$. An easy way to determine if continuous variables are independent is to first check that the domain only contains constants, is rectangular, and that the joint pdf can be written as a product of a function of X only and a function of Y only (see below for more explicit if-then statement)

Conditional expectation

Expected value taken based on a certain condition called a realization. Assuming Y is the realization for a random variable X,

$$E[X] = E[E[X|Y]]$$

Let there be k many values that Y can take that are mutually exclusive and exhaustive. Then

$$E[X] = E\big[E[X|Y]\big] = \sum_{n=1}^{k} P(Y=y_n)\cdot E[X|Y=y_n]$$

Note that $E(X+Y) = E(X) + E(Y)$.

Note that $E(XY) = E(X)E(Y)$ is not necessarily true. It is true when X and Y are independent.

Variance expressions originate from that variance of a random variable is the expected value of the squared difference from its mean. So

$$Var(XY) = E\left[\left(XY - E(XY)\right)^2\right] = E\left[\left(XY - \frac{8}{9}\right)^2\right]$$

Remember that if a and b are constants,

$$E(aX+b) = aE(X) + b$$

and

$$Var(aX+b) = a^2Var(X).$$

Similarly, if a, b, c, and d are all constants

$$Cov(aX + b, cY + d) = acCov(X, Y)$$

Independence Check

If joint probability function $f(x, y)$ can be separated into function of x times some function of y, meaning

$$f(x, y) = g(x)h(y)$$

and the domain of $f(x, y)$ only contains constants, then x and y are independent.

Chapter 16 – Transformations

Transformation of discrete random variables

Let X be a *discrete* random variable and let g be a function. The variable $Y = g(X)$ is a discrete random variable with pmf

$$f_Y(y) = P(Y = y) = \sum_{\forall x | g(x) = y} f_X(x)$$

Cdf method

This is used for the transformation of *continuous* random variables. The idea is to find the cdf of the new random variable and then find the pdf using the Fundamental Theorem of Calculus. Suppose X is a continuous random variable with cdf $F_X(x)$. Let $Y = g(X)$. We can find the cdf of Y as

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P\big(X \leq g^{-1}(y)\big) = F_X(g^{-1}(y))$$

To get the pdf of Y we need to take the derivative of the cdf.

Pdf method

We can find the pdf of Y by differentiating the cdf

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\big(g^{-1}(y)\big) = f_X\big(g^{-1}(y)\big) \left| \frac{d}{dy} g^{-1}(y) \right|$$

So long as g^{-1} is differentiable, we can use this method to directly obtain the pdf of Y.

Conversely, we can find the cdf by integrating the pdf.

Chapter 17 – Estimation Methods

Statistical models	Describe one or more variables and their relationships; used to make decisions about a population, predict future outcomes
Inferential statistics	Used to draw conclusions about an underlying process
Probability vs. statistics	In probability, we describe what we expect to happen if we know the underlying process; in statistics, we don't know the underlying process, and must infer based on representative samples.
Estimation	Used to determine population parameters when given a sample; estimates the parameters of the probability distribution a given sample is assumed to follow
Population moments	
First moment (mean)	$E(X) = \mu$
kth central moment	$E(X^k)$
k moment around the mean	$E[(X-\mu)^k]$
2nd moment around the mean (variance)	$E[(X-\mu)^2]$
Sample moments	Suppose $X_1, X_2, \dots, X_n$ is a sequence of independent, identically distributed (iid) random variables with some distribution parameters θ.
kth central sample moment	$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$
kth sample moment around the mean	$\hat{\mu}'_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^k$
Method of moments	Let X be a random variable with some type of distribution that has parameter θ. We calculate $E(X)$ in terms of θ and set this equal to the 1st sample moment (mean) to solve for θ.

Unbiased

A property causing the estimate for the square root of standard deviation is different from sample standard deviation. If taking the expected a random variable estimator equals the parameter being estimated, then the estimator is unbiased. This is written mathematically as

$$E(\hat{\theta}) = \theta$$

In other words, unbiased means on average, the estimator will equal the true value. It is not a necessary quality for estimations but is desirable.

Maximum likelihood

Finding values of parameters that would make the observed data most likely.

Likelihood function

Suppose $x_1, x_2, \dots, x_n$ is an iid random sample from a distribution with mass/ density function $f_X(x; \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ are the parameters. The symbol $\boldsymbol{\theta}$ is a vector.

The likelihood function is denoted as $L(\boldsymbol{\theta}; x_1, x_2, \dots, x_n) = L(\boldsymbol{\theta}; \boldsymbol{x})$ which, since our sample is iid, we can write as a product of pmfs/ pdfs according to

$$L(\theta; x) = \prod_{i=1}^{n} f_X(x_i; \boldsymbol{\theta})$$

This function is really the pmf/pdf except instead of the variables being random and the parameters being fixed, the values of the variables are known and the parameters are unknown.

Log likelihood function

$$\ell(\boldsymbol{\theta}, \boldsymbol{x}) = logL(\boldsymbol{\theta}; \boldsymbol{x}) = \sum_{i=1}^{n} \ln\left(f_X(x_i; \boldsymbol{\theta})\right)$$

We maximize this function to maximize the likelihood function. This is easier because we do not have to use the product rule since the log of a product is a sum. Maximizing the likelihood function with respect to the parameters gives the parameter values that make the observed values the most likely.

Chapter 18 – Statistical Modeling Case Study
Chapter 19 – Hypothesis Testing
Chapter 20 – Empirical $\boldsymbol{p}$-values

Point estimate	Difference between groups of a sample. For example, if we are interested in exploring differences between promotions at a bank for males and females, and we conduct an experiment where 58.3% of females are promoted while 87.5% of males are promoted, our point estimate is 87.5-58.3=29.2%
Hypothesis test	A statistical technique used to evaluate competing claims using data
$\boldsymbol{H_0}$ **null hypothesis**	The first hypothesis, often is skeptical of research claim
$\boldsymbol{H_A}$ **alternative hypothesis**	A hypothesis different from the null hypothesis
Test statistic	A single value metric designed to measure the relationship between variables
$\boldsymbol{p}$**-value**	The probability of something as or more extreme than what is observed assuming the null hypothesis is true
Statistically significant	When the p-value is small, meaning less than a previously set threshold
Significance level	The threshold value that determines statistical significance; represented by α and usually set to 0.05, $P(type\ 1\ error) = \alpha$
Decision Errors	
Type 1 error, false positive	Rejecting the null hypothesis when it is actually true
Type 2 error, false negative	Failing to reject the null hypothesis when the alternative is actually true
Confirmation bias	Looking for data that supports ideas; only aiming to test hypothesis that agree with our already established beliefs
One-sided hypothesis	Hypothesis that explore only one direction of possibilities; for parameter θ such hypotheses look like $H_0{:}\,\theta = \theta_0\ H_1{:}\,\theta > \theta_0$ or $H_0{:}\,\theta = \theta_0\ H_1{:}\,\theta < \theta_0$

Two-sided hypothesis	Hypothesis that consider all possibilities; for parameter θ such hypotheses look like $H_0: \theta = \theta_0 \; H_1: \theta \neq \theta_0$
Hypothesis testing using probability models	A class of hypothesis testing where the null hypothesis specifies a probability model
Power	The probability of rejecting the null hypothesis when the alternative hypothesis is true
Permutation test	Shuffling the data grouping to simulate resampling under an assumed null hypothesis of a certain variable being independent of the grouping

Steps to Determine Validity of Research Question

1. State null and alternative hypothesis
   - This involves framing the research question in terms of two hypotheses, the null and alternative hypotheses
   - Typically, the null case, similar to a presumption of innocence, assumes the relationship the research question aims to find is non existent
   - The alternative typically is the claim of the research question
2. Compute a test statistic
   - This involves choosing a metric that allows the stated hypotheses to be more specific
   - This metric should be calculated for the observed values
3. Determine the p-value
   - Use randomization to determine the p-value
   - Include the observed data in the calculation of the p-value; this involves adding 1 to both the numerator and denominator of the probability of the p-value
4. Draw a conclusion
   - Determine presence or lack of statistical significance
   - Note, insufficient evidence of the alternative hypothesis does not necessarily prove the null hypothesis, it is merely enough to refute adoption of the alternative hypothesis, much like how insufficient evidence of guilt is not proof of innocence, but still prevents a guilty verdict.

Choosing a Significance Level

If we want to minimize or caution type-1 errors, we choose a low significance level. If we want to minimize or caution type-2 errors, we choose a high significance level. Of course, minimizing one error makes the other more likely.

How to Use a Hypothesis Test

1. Frame the research question in terms of hypotheses.
   Hypothesis tests are appropriate for research questions that can be summarized in two competing hypotheses. The null hypothesis H_0 usually represents a skeptical perspective or a perspective of no difference. The alternative hypothesis H_A usually represents a new view or a difference.

2. Collect data with an observational study or experiment.
   If a research question can be formed into two hypotheses, collect data to run a hypothesis test. If the research question focuses on associations between variables but does not concern causation, run an observational study. If the research question seeks a causal connection between two or more variables, then an experiment should be used.

3. Analyze the data.
   Choose an analysis technique appropriate for the data and identify the p-value.

4. Form a conclusion.
   Using the p-value from the analysis, determine whether the data provide statistically significant evidence against the null hypothesis. Also, be sure to write the conclusion in plain language so casual readers can understand the results.

Chapter 21 – Central Limit Theorem

Central Limit Theorem

Let $X_1, X_2, \dots, X_n$ be a sequence of iid, independent and identically distributed, random variables from a distribution with mean μ and standard deviation $\sigma < \infty$ (finite variance). Then

$$\bar{X} \sim Norm\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Where $\sim$ signifies "approx." In words, this is saying the distribution the sample means across multiple simulations of the iid sequence should follow a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Rearranging,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim Norm(0,1)$$

Chi-squared Lemma

Let $S^2 = \frac{\sum(X_i - \bar{X})^2}{n-1}$. Further let $X_1, X_2, \dots, X_n$ be an iid sequence of random variables from a normal population with mean μ and standard deviation σ. Then

$$\frac{(n-1)S^2}{\sigma^2} \sim Chisq(n-1)$$

***t* Lemma**

Let $X_1, X_2, \dots, X_n$ be an iid sequence of random variables, each with mean μ and sample deviation S. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

This expression is referred to as the t statistic and can be used as a test statistic in hypothesis testing, just like mean or variance etc.

Degrees of freedom

The number of independent pieces of information that go into calculating the estimate; number of values free to vary in a data set; if n is the sample size, then $df = n - 1$;

Ex. If we have a list of n values $x_1, x_2, \dots, x_n$ and their mean must equal a fixed μ according to

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \mu$$

Then we have $n-1$ degrees of freedom as $n-1$ x values can be anything while the last x value will have to be some specific value to make the equation true

Standard error	Standard deviation over the square root of the sample size; $SE = \frac{S}{\sqrt{n}}$

Summary and Rules of Thumb

1. The central limit works regardless of the distribution. However, if the parent population is highly skewed, then more data is needed. The CLT works well once the sample sizes exceed 30 to 40. If the data is fairly symmetric, then less data is needed.
2. When estimating the mean and standard error from a sample of numerical data, the t distribution is a little more accurate than the normal model. But there is an assumption that the parent population is normally distributed. This distribution works well even for small samples as long as the data is close to symmetrical and unimodal.
3. For medium samples, at least 15 data points, the t distribution still works as long as the data is roughly symmetric.
4. For large data sets 30-40 or more, the t or even the normal can be used.

Assumptions of t Distribution

1. Independence of observations
2. Observations come from a nearly normal distribution. Moderate skew is acceptable when sample size is 30 or more, large skew acceptable when size is about 60 or more.

Chapter 22 – Confidence Intervals

Range of values	Possible values for a parameter
Confidence interval	A plausible range of values for a population parameter; a range of values such that with p probability, the range will contain the true unknown value of the parameter
Point estimate	Best guess for value of parameter
Two-sided intervals	Bounding the interval on both sides
One-sided interval	Bounding the interval on one side
Standard error in difference of sample proportions	$SE_{\hat{p}_1-\hat{p}_2} = \sqrt{SE_{\hat{p}_1}^2 + SE_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ Standard error is standard deviation but instead for a statistic rather than a variable
Success-failure condition	For a distribution to be approximated as normal, there must be at least 10 successes (s) and 10 failures (f); when $s \geq 10$ and $f < 10$, the data is left skewed; when $s < 10$ and $f \geq 10$, the data is right skewed
Margin of error	Half the width of the confidence interval; also the difference between the observed and estimated value

Building Confidence Intervals

1. Identify the parameter that is to be estimated
2. Identify a good estimate for that parameter
3. Determine the distribution of the estimate or a function of the estimate
4. Use this distribution to obtain a range of feasible values (confidence interval) for the parameter (If μ is the parameter of interest, using the CLT, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim Norm(0,1)$ which can be solved for μ to obtain a range $\mu \in (\bar{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$

Hypothesis Test for Two Proportions Conditions

The difference between two proportions $\hat{p}_1 - \hat{p}_2$ tends to follow a normal distribution when

1. Each proportion separately follows a normal model
2. The two samples are independent of each other

A Note on Interpreting Confidence Intervals

Stating “we are $x\%$ confident that the population parameter is between…” should not be interpreted as “there is $.x$ probability that our parameter is between…”; rather, it should be interpreted as “we have $x\%$ confidence that our true parameter is plausibly between…”.

If I run the test y many times, I expect $y*.x$ of the tests to produce an interval with the true value inside.

Chapter 23 – Bootstrap

Bootstrapping

Allows us to simulate sampling distribution by resampling from the sample; involves resampling with replacement from original sample, calculating and recording sample statistic for bootstrapped sample, then repeating process many times; way to find variability in estimate

Resampling

Sampling from original sample with replacement

Plug-in principle

If something is unknown, substitute an estimate of it

Pooled standard deviation

Using data from both samples to better estimate the standard deviation and standard error; used when standard deviations of two groups are near equal; gives better estimate of standard deviation from each group and allows larger degrees of freedom for t distribution; calculated by

$$s_{pooled}^2 = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

Where n is size of respective group and s is standard deviation of respective group

Proportion test

Tests whether proportion of k successes of n conditions is truly p; null hypothesis is that it is

Ex: If it is desired to test they hypothesis that one-half of the days in a year are sunny ($p = .5, n = 365$) but 300 days were sunny, ($k = 300$) we would run a proportion test

Chapter 24 – Additional Hypothesis Tests

Contingency table

Case mapping of two variables; Ex:

	Heads	Tails
Day	22	18
Night	17	23

Pearson chi-squared test statistic

Let e_{ij} be the expected count in the $i$the row and jth column in a contingency table with r rows and c columns under the null hypothesis. The test statistic is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(n_{ij}-e_{ij}\right)^2}{e_{ij}}$$

To find e_{ij}, recognize that under the null hypothesis (independence), the joint probability is equal to the product of the two marginal probabilities. Essentially compares observed counts with expected counts under null hypothesis.

The chi-squared distribution is skewed right and bounded on the left at zero since it is the sum of squared differences. The degrees of freedom are $df = (r-1)(c-1)$

More generally, we have

$$\chi^2 = \sum_{All\ i}\left(\frac{Exp_i - Obs_i}{Exp_i}\right)^2$$

$$df = \#\ of\ bins - 1 - \#\ of\ estimated\ parameters$$

Analysis of variance (ANOVA) with F distribution

Uses a single hypothesis test to check whether the means across many groups are equal. They hypotheses are:

H_0: The mean outcome is the same across all k groups; notationally, $\mu_1 = \mu_2 = \cdots = \mu_k$where μ_i is the mean of group i

H_A: At least one mean is different

Answers question "is the variability in the sample means so large that it seems unlikely to be from chance alone?"

Data snooping/ data fishing

Examining all data beforehand to decide which parts to formally test; leads to an inflation in Type 1 error

Mean square between groups variability (MSG)	Scaled variance formula for means. Has degrees of freedom $df_G = k - 1$ for k groups. Let $\bar{x}$ represent the mean of outcomes across all groups. Then the MSG is computed as $MSG = \frac{1}{df_G} SSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i (\bar{x}_i - \bar{x})^2$ Where SSG is called the sum of squares between groups and n_i is the sample size of group i.
Mean square error (MSE)	A pooled variance estimate which measures variability within groups. Has degrees of freedom $df_E = N - k$ where $N = \sum n_i$. Is standard form of SSE $MSE = \frac{1}{df_E} SSE$
Sum of squares total (SST)	$\sum_{i=1}^{n} (x_i - \bar{x})^2$
Sum of squared error (SSE)	$SSE = SSET - SSG$ $= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$ Where s_i^2 is the sample variance of the residuals in group i
***F* Test**	Uses F statistic to evaluate hypothesis $F = \frac{MSG}{MSE}$ An F distribution has to associated parameters: df_1 and df_2. For the F statistic in ANOVA $df_1 = df_G$ and $df_2 = df_E$. Larger values of F represent stronger evidence against the null hypothesis.

Conditions to Check Before ANOVA and *F* Distribution

i. The observations are independent within and across all groups
ii. The data within each group is nearly normal, and
iii. The variability across the groups is about equal

Chapter 25 – Predictive Statistical Modeling Case Study

Residual	Differences between the observed values and the values predicted by the line $e_i = y_i - \hat{y}_i$
Linear regression	Finding a line that best fits the data; usually done my finding a line that minimizes the sum of squared residuals
Extrapolation	When predictions are made for values of x (according to linear regression line) that are beyond the range of the observed data
Multiple $\boldsymbol{R}$ squared ($\boldsymbol{R^2}$)	Represents the proportion of variability in the response variable that is explained by the explanatory variable

Chapter 26 – Linear Regression Basics

Linear model

Estimates relationship between a response variable and its predictor variable(s) to both predict values and infer relationships

Linear regression model

Suppose we are interested in exploring the relationship between one response variable Y and n predictor variables X. We postulate the linear relationship between the two as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

When $n = 1$ we call this a **simple linear regression** model. Then β terms are referred to as coefficients

Residual, error term

The deviation from the linear fit; assumed to follow a normal distribution with mean 0 and a constant standard deviation

Method of least squares

Finding the parameter values that minimize the squared vertical distance between the points and the resulting line; in other words, minimizing the squared residuals

This method is used for three following (somewhat circular) reasons

i. It is the most common
ii. Computing a line based on least squares was much easier by hand before computers were available
iii. A residual twice as large as another residual is more than twice as bad and squaring the results accounts for this

$\boldsymbol{\beta_0}$

Represents the average value of the response when the predictor is zero; known as

$\boldsymbol{\beta_1}$

Represents the average increase in the response variable per unit increase in the predictor variable; known as slope coefficient

Average

Expected response of model for given input; mathematically is

$$E(Y|X = x) = E(\beta_0 + \beta_1 x + e) = E(\beta_0 + \beta_1 x) + E(e) = \beta_0 + \beta_1 x$$

In general, we have $Data = Fit + Residual$

Calculations for Method of Least Squares

Fit line will take on formula $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 x_1$. This implies it will always go through the point $(\bar{x}, \bar{y})$

A formula for slope of the regression line that links it with correlation is

$$\hat{\beta}_1 = \frac{s_y}{s_x} R$$

Where R is correlation and s is the respective sample standard deviations of the explanatory and response variables.

The standard deviation of the error can be found by

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \widehat{e_2^{i}}}$$

where $\hat{e}_i$ is the observed ith residual ($\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1$).

Linear Regression Assumptions

To assess whether a linear model is reliable, we must check for

1. Linearity of observations
2. Independence of observations
3. Nearly normal distribution of residuals
4. Constant variability (error variance)

The last three assumptions are not necessary for estimating the relationship but rather for inferring it.

Chapter 27 – Linear Regression Inference

Average or expected value given predictor $\boldsymbol{x_*}$	The resulting prediction of the model give input x_* is $\hat{Y}_* = \widehat{\beta_0} + \widehat{\beta_1} x_*$
Estimators	What $\widehat{Y_*}, \widehat{\beta_0}$ and $\widehat{\beta_1}$ are after modeling a random sample
Prediction Interval	Building an interval for a predicated observation Y_{new} at x_* is according to $Y_* \in \left(\hat{Y}_* \pm t_{\frac{\alpha}{2}, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right)$ Prediction intervals are wider since they are intervals on individual observations and not on averages.

Under the assumption that the error term is normally distributed, the distributions of our estimators behave according to

$$\varepsilon = N(0, \sigma)$$

$$\widehat{\beta_0} \sim N\left(\beta_0, SE = \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}} \right)$$

$$\widehat{\beta_1} \sim N\left(\beta_1, SE = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}} \right)$$

σ is residual standard error

$$\hat{Y}_* \sim N\left(\beta_0 + \beta_1 x_*, SE = \sigma \sqrt{\frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right)$$

Ex. Preforming Inference on $\boldsymbol{\widehat{\beta_1}}$

By above, the $\widehat{\beta_1}$ standard normal distribution transformation is according to

$$\frac{\widehat{\beta_1} - \beta_1}{\frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}} \sim N(0,1)$$

To estimate the error standard deviation σ, we have

$$\hat{\sigma} = \sqrt{\frac{1}{n-2} n \sum_{i=1}^{n} \widehat{e_i^2}}$$

Where $\hat{e}_i$ is the observed ith residual ($\widehat{e_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$). Since we replaced the population standard deviation with an estimation, the resulting random variable follows a t distribution according to

$$\frac{\widehat{\beta_1} - \beta_1}{\frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}} \sim t(n-2)$$

Where we have $n - 2$ degrees of freedom because we had to estimate two parameters. To get a confidence interval for β_1 we have

$$\beta_1 \in \left(\widehat{\beta_1} \pm t_{\frac{\alpha}{2}, n-2} \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}} \right)$$

We can evaluate the null hypothesis $H_0: \beta_1 = \beta_1^*$ using the t statistic having a $t(n-2)$ distribution as our test. The t statistic is calculated as

$$t = \frac{\hat{\beta}_1 - \beta_1}{SE} \sim t(n-2)$$

Chapter 28 – Regression Diagnostics

R-squared, coefficient of determination

Correlation squared; one measure of goodness of fit; ratio of variance in response explained by the model to overall variance of the response; Composed by

$\sum_{i=1}^{n}(y_i - \bar{y})^2 = SS_{Total}$

$\sum_{i=1}^{n}(\widehat{y_i} - \bar{y})^2 = SS_{Regression}$ (variation due to linear relationship between y and the predictor variables)

$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{Error}$ (variation due to random scatter)

$$R^2 = SS_{Regression}/SS_{Model}$$

Proportion of overall variation in the response that is explained by the linear model; can be between 0 and 1 where high values indicate tight fit, low indicate large scatter

F-statistic

Given by

$$\frac{n-p-1}{p}\frac{\sum(\hat{y}_i - \bar{y})^2}{\sum e_i^2}$$

Under a null hypothesis that all non-intercept coefficients are zero; stated mathematically as

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_n = 0$$

This model follows an F distribution with parameters p and $p - 1$

Homoskedastic

Errors have constant variance

Five plot(lm()) plots

Residuals vs fitted

Assesses: linearity of model and homoscedasticity (constant variance);

Ideally: red line should coincide with dashed horizontal line and the residuals (points) should be centered around the dashed line

Normal Q-Q

Assesses: normality of residuals

Ideally: Points are on line, which means distribution is normal

	If right skewed, right will be above the line; left tail will level off which indicates concentration of points on the left If left skewed, left will be below the line; right tail will level off which indicates concentration of points on the left
Scale –location plot	Plot of fitted values versus square root of absolute value of standardized residuals; is residual divided by its standard deviation according to $$e_i' = \frac{e_i}{s}$$ Assesses: non-constant error variance; horizontal red line indicates error variance Ideally: Horizontal red line is constant
Outliers and leverage	Two types of outliers; outliers in the response and outliers in the explanatory
Outliers	Outlier in response variable
Leverage points	Outlier in explanatory variable
Influential point	An outlier that drastically alters the regression output; typically a high leverage point
High leverage points	Points that fall horizontally away from the center of the “cloud” that the point values make; can strongly influence the slope of the least squares line
Residuals vs leverage plot	Identifies: Influential observations which are values with high Cook’s distance Ideally: red line does not cross Cook’s distance topographic lines

Regression Diagnostics

1. Issues with following assumptions about the error structure:
    a. Errors are normally distributed
    b. Errors are independent
    c. Errors have variance

2. Issues with following assumption about fit:
    a. Assume model is correct
3. Problems with outliers/ leverage points
    a. Outliers can give mistaken information about model's fit
4. Missing predictors

Chapter 29 – Simulation Based Linear Regression

Residual resampling	New data set has all of the predictor values from the original data set and a new response is created by adding a bootstrap resampled residual to the fitted function. To summarize: Suppose there are n observations, each with response Y and some number of explanatory X's. We can bootstrap by a. Bootstrap observations b. Bootstrap residuals
Categorical predictor	A predictor variable that is a categorical variable
Contrasts	Coding of covariate into dummy variables
Reference category	When all dummy variables equal zero
Pairwise comparisons	Allows us to conduct hypothesis tests of certain categorical comparisons

Process for Bootstrapping Residuals

1. Fit the regression model to original data
2. Compute the predicted response values $\hat{Y}_i$ and residuals $e_i = Y_i - \hat{Y}_i$
3. Create a bootstrap sample using the same X explanatory values as the original data but with new Y response variables notated as $Y_i^* = \hat{Y}_i + e_i^*$ where e_i^* is sampled randomly with replacement from the original residuals

Contrasts

In the case of a categorical covariate with k levels where $k > 2$, we must include $k - 1$ dummy variables in the model. Each dummy variable takes on a value of 0 or 1.

> Ex. If a covariate has $k = 3$ categories or levels (A, B, C), we create two dummy variables X_1 and X_2 which can only take values of 0 or 1. If $X_1 = 1$ then the covariate takes the value A. Likewise, if $X_2 = 1$, then we state that the covariate takes the value B. If we have the reference category, the covariate takes the value C.
>
> The linear model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$. We have
>
> $$E(Y|X = A) = \beta_0 + \beta_1$$
>
> $$E(Y|X = B) = \beta_0 + \beta_2$$
>
> $$E(Y|X = C) = \beta_0$$

Reference category is factored into the intercept, other factor categories "amend" the intercept

Chapter 30 – Multiple Linear Regression

Multiple regression	Extends simple two-variable regression to the case that still has one response but many predictors (denoted as x_1, x_2, x_3 …)
Collinear	Describes two correlated predictor variables
Adjusted R^2	Adjusted R^2 as a tool for model assessment with multiple variables is computed as $R^2 = 1 - \frac{sum\ of\ squares\ of \frac{residuals}{n-k-1}}{sum\ of\ squares\ of\ the \frac{outcome}{n-1}}$ Where n is the number of cases used to fit the model and k is the number of predictor variables

β_1 for a certain variable is interpreted the same as the simple case yet with the only difference that it is interpreted as such when all other explanatory variables are held constant

Chapter 31 – Logistic Regression

Logistic regression

A type of generalized linear model for building models when there is a categorical response variable Y_i with two levels

Model relates probability of a success condition $\widehat{p_i}$ for a response variable Y_i to values of predictors according to

$$trans(p_i) = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i}$$

For k predictor variables of an observation i

Generalized linear model

Used for response variables where the assumptions of normally distributed errors is not appropriate; two-stage modeling approach where response variable is modeled using a probability distribution and the parameter distribution is modeled using predictors and multiple regression

Fisher Exact Test

Hypothesis permutation test using a hypergeometric distribution; null hypothesis is true odds ratio of predictors is equal to one

Logit transformation

The transformation we select to use in logistic regression

$$logit(p_i) = \ln\left(\frac{p_i}{1 - p_i}\right)$$

Setting this equal to the model in the logistic regression cell, we solve for p_i (the probability of the default case) and see

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i})}}$$

True odds ratio

$$\frac{p_{response|predictor\ condition\ satisfied}}{p_{response|predictor\ condition\ not\ satisfied}}$$

Residual deviance

Model is not fit using least squares, but rather using deviance, which is twice the negative of the log likelihood

Confusion matrix

A 2x2 matrix comparing the predictions of our model to the actual outcomes; assess how good a model is

To translate probability into model prediction, generally, when $\hat{p}_i > 0.5$, the model predicts a success; otherwise, model predicts failure. Matrix looks like

		Actual	
		0	1
Prediction	0	$n_{0,0}$	$n_{0,1}$
	1	$n_{1,0}$	$n_{1,1}$

Accuracy	Metric of confusion matrix that determines how good model is; calculated by $\frac{n_{0,0}+n_{1,1}}{n_{0,0}+n_{1,1}+n_{0,1}+n_{1,0}}$

Assumptions for Logistic Regression

1. Each predictor x_i is linearly related to $logit(p_i)$ if all other predictors are held constant
2. Each outcome Y_i is independent of the other outcomes
3. There are no influential data points
4. Multicollineraity is minimal

Math 378 – Applied Statistical Modeling

A Note on Notational Use for this Section

- n is the number of distinct data points
- p is the number of available predictors used to create models
- k is the number of classes of a response categorical variable
- y represents a response variable, x represents a predictor

Chapter 1 - Introduction

The Supervised Learning Problem

- **Outcome measurement** Y (also called dependent variable, response, target)
- Vector of p **predictor measurements** X (also called inputs, regressors, covariates, features, independent variables)
- In the **regression problem**, Y is quantitative
- In the **classification problem**, Y takes values in a finite, unordered set
- We have **training data** $(x_1, y_1), \dots, (x_N, y_N)$ that are observations (also called examples, instances) of these measurements to help fit, train, or tech the model
- Objectives
  - Accurately predict unseen test cases
  - Understand which inputs affect the outcome, and how
  - Assess the quality of our predictions and inferences

Unsupervised Learning

- No outcome variable, just a set of predictors (features) measured on a set of samples
- Objective is more fuzzy – find groups of samples that behaves similarly, find features that behave similarly, find linear combinations of features with the most variation
- Difficult to know how well you are doing
- Different from supervised learning, but can be useful as a pre-processing step for supervised learning
- In the **clustering problem**, we attempt to ascertain on the basis of certain variables, whether observations fall into distinct groups

Statistical Learning vs Machine Learning

- Machine learning arose as a subfield of Artificial Intelligence
- Statistical learning arose as a subfield of Statistics
- There is much overlap – both fields focus on supervised and unsupervised problems

  - Machine learning has a greater emphasis on large scale applications and prediction accuracy
  - Statistical learning emphasizes modes and their interpretability, and precision and uncertainty
- The distinction has become more and more blurred and there is a great deal of “cross-fertilization”
- Machine learning has the upper hand in marketing

Chapter 2 – Statistical Learning

What is $f(X)$ used for?

- With a good f we can make **predictions** of Y at new points with $X = x$. Here, we can treat f as a **black box** where we do not necessarily need to understand the exact form of f
- We can understand which components of X are important in explaining Y and which are irrelevant
- Depending on the complexity of f, we may be able to understand how each component of X_j affects Y. This is known as **inference**. Here, we can not treat f as a black box since we must know its form. We are interested in answering
    - Which predictors are associated with the response?
    - What is the relationship between the response and each predictor?
    - Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

The regression function $f(x)$

- $f(x) = E(Y|X = x)$ defines the regression function
- Is the ideal or optimal predictor of Y with regard to mean-squared prediction error; this function minimizes $E\left[\left(Y - g(X)\right)^2|X = x\right]$ over all functions g at all points $X = x$
- $\epsilon = Y - f(x)$ is the irreducible error – i.e. even if we knew $f(x)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values
- For any estimate $\hat{f}(x)$ of $f(x)$, we have
  $E\left[\left(Y - \hat{f}(X)\right)^2 |X = x\right] = Reducible + Irreducible = \left[f(x) - \hat{f}(x)\right]^2 + Var(\epsilon)$

How to estimate f

- Typically we have few if any data points with $X = x$ exactly
- So we cannot compute $E(Y|X = x)$
- Relax the definition and let $\hat{f}\ (x) = Ave(Y|X \in \mathcal{N}(x))$ where $\mathcal{N}(x)$ is some neighborhood of x

Nearest Neighbor Issues

- Nearest neighbor averaging can be pretty good for small p and large-ish N
- Nearest neighborhood methods can be lousy when p is large. This is due to the curse of dimensionality, which means that nearest neighbors tend to be far away in high dimensions

Parametric and Structured Models

- We first make an assumption about the functional form, or shape of f. Next, we procure training data to fit or train the model
- The linear model is an important example of a parametric model
$$f_L(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$
- A linear model is specified in terms of $p + 1$ parameters $\beta_0, \beta_1, \dots, \beta_p$

- We estimate the parameters by fitting the model to training data, typically for a linear model using the **least squares**
- **Parametric** models reduce the problem of estimating f down to a problem of estimating a set of parameters
- Although it is almost never correct, a linear model often serves as a good and interpretable approximation to the unknown true function $f(X)$
    - This problem can be addressed by choosing a **flexible** method that can fit many different possible functional forms for f
    - However, more flexible models can lead to model **overfitting** the data, a phenomena that occurs when models follows deviations or **noise** too closely

Non-Parametric Methods

- Such methods do not make explicit assumptions about the functional form of f, they instead seek an estimate of f that gets as close to the data points as possible and smoothly
- These approaches can fit a wider range of possible shapes for f, however, much more data is needed to obtain accurate estimates of f

Mean Squared Error (MSE)

- The most commonly used measure of quality of fit is the least squares method
- This aims to minimize mean squared error of the which is given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  Where $\hat{f}(x_i)$ is the prediction given by $\hat{f}$ for the ith observation
- MSE is initially calculated using the training data, but we really want to minimize the MSE of $\hat{f}$ using test data not previously used to train the model. However, when test data is unavailable, it is not necessarily true that a low MSE of training data corresponds to a low MSE of test data
- If there is a low training MSE but a high test MSE then our model is overfitting. Also, in the case where a less flexible model would have yielded a smaller test MSE, our model is overfitting.

Some Modeling trade-offs

- Prediction accuracy vs. interpretability
    - Restrictive models are more interpretable for inference than more flexible models
    - Sometimes less flexible methods are better for prediction because more flexible methods will cause overfitting
- Good fit vs. over-fit or under-fit
- Parsimony vs. black-box,

Bias-Variance Trade-off

Suppose we have fit a model $\hat{f}(x)$ to some training data Tr, and let (x_0, y_0) be a test observation drawn from the population. If the true model is $Y = f(X) + \epsilon$ (with $f(x) = E(Y|X = x)$), then

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = Var\left(\hat{f}(x_0)\right) + [Bias\left(\hat{f}(x_0)\right)^2 + Var(\epsilon)$$

The expectation averages over the variability of y_o as well as the variability in Tr. Note that $Bias(\hat{f}(x_0)) = E\left[\hat{f}(x_0)\right] - f(x_0)$.

- The term **variance** refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Generally, more flexible models have higher variance. Ideally, we wish to minimize variance.
- The term **bias** refers to the error that is introduced by approximating a real-life scenario with a model inevitably simpler than real-life complexity. Generally, more flexible models have lower bias. Ideally, we wish to minimize bias.
- The relative rate of change of these two quantities determines whether the test MSE increases or decreases.
    - As we increase the flexibility of a class of methods, the bias tends to initially decrease faster than the variance increases. Consequently, the expected test MSE declines.
    - However, at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance. When this happens the test MSE increases.
- Hence, choosing flexibility amounts to a bias-variance trade-off

Classification Problems

Here the response variable Y is qualitative. Our goals are to

- Build a classifier $C(X)$ that assigns a class label from C to a future unlabeled observation X.
- Assess the uncertainty in each classification
- Understand the roles of the different predictors among $X = (X_1, X_2, \dots, X_p)$

Is there an ideal $C(X)$? Suppose the K elements in C are numbered 1,2, … K. Let

$$p_k(x) = \Pr(Y = k|X = x)\,, k = 1,2, \dots, K$$

These are the conditional class probabilities at x. Then the Bayes optimal classifier at x is

$$C(x) = j \ if \ p_j(x) = max\{p_1(x), p_2(x), \dots, p_K(x)\}$$

Nearest-neighbor averaging can be used as before, but it also begins to fail as the dimension grows. However, the impact on $\hat{C}(x)$ is less than on $\hat{p}_k(x), k = 1, \dots, K$

- Typically we measure the performance of $\hat{C}(x)$ using the misclassification **error rate**:

  $$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

  Where $\hat{y}_i$ is the predicted class lable for the ith observation using $\hat{f}$ and $I(y_i \neq \hat{y}_i)$ is an **indicator variable** that equals 1 if the prediction is wrong and zero if it is correct

- The **Bayes classifier** assigns each observation to the most likely class given its predictor values.
  - Stated mathematically, this means we assign a test observation with predictor vector x_o to the class j for which

    $$\Pr\,(Y = j|X = x_0)$$

    is largest.
  - In an example where there are two classes, the Bayes classifier will predict class one if $\Pr(Y = 1|X = x_0) > 0.5$ and class two otherwise. The Bayes decision boundary occurs when the predictor vector yields a probability that is exactly 50%
  - The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate. This error rate will be

    $$1 - E(\max_j \Pr\,(Y = j|X))$$
- The K**-nearest neighbors (KNN)** classifier is classifies a given observation based on the highest estimated probability.
  - The KNN takes the following steps
    - Given a natural number K and a test observation x_0, the KNN classifier identifies the K nearest points in the training data closest to x_0 represented by $\mathcal{N}_0$.
    - It then estimates the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j. Stated mathematically,

      $$\Pr(Y = j|X = x_0) = \frac{1}{K}\sum_{i \in \mathcal{N}_0} I(y_i = j)$$
    - Lastly, KNN classifies the test observation x_0 to the class with the largest probability
  - A low K value high flexibility, where as K grows, the method becomes less flexible

Chapter 3 – Linear Regression

Simple Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \dots, X_p$ is linear
- True regression functions are never linear
- Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically
- Simple Linear Regression Using a Single Predictor X
  - We assume a model

    $$Y = \beta_o + \beta_1 X + \epsilon$$

    Where β_0 and β_1 are two unknown constants that represent the **intercept** and **slope**, also known as **coefficients** or **parameters**, and ϵ is the error term.

    Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients that are calculated using training data, we predict future outcomes using

    $$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$$

    Where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The hat denotes an estimates value. We sometimes describe the above model as saying we are regressing Y on X.
- Estimation of the Parameters by Least Squares
  - Let $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith residual value of X. Then $e_i = y_i - \hat{y}_i$ represents the $i$the **residual**
  - We define the **residual sum of squares** (RSS) as

    $$RSS = e_1^2 + e_2^2 + \cdots + e_n^2,$$

    or equivalently as

    $$RSS = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$
  - The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Thus minimizing values can be shown to be

    $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

    $$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

    Where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \sum_{i=1}^{n} x_i$ are the sample means
- Assessing the Accuracy of the Coefficient Estimates
  - We may want to estimate parameters of the larger population by using data provided by the model
    - Typically, such methods are **unbiased** in the sense that if we could average a huge number of estimates of the parameter based on a huge number of sample sets from the larger population, we would have the exact value of the population parameter

        - Unbiased holds for estimating the mean and the least squared coefficients ($\beta_0$and β_1)
    - The standard error of an estimator reflects how it varies under repeated sampling. We have
$$SE(\hat{\mu})^2 = \frac{\sigma^2}{n}, \qquad SE\left(\hat{\beta}_0\right)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \qquad SE\left(\hat{\beta}_1\right)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
where $\sigma^2 = Var(\epsilon)$.
    - Typically, σ is not known but can be estimated from the data according to the **residual standard error** which is given by $RSE = \sqrt{\frac{RSS}{n-2}}$
    - These standard errors can be used to compute confidence intervals; for example, a confidence interval that has a 95% chance of containing the true value of β_1 takes the form
$$\hat{\beta}_1 \pm 2 \cdot SE(\hat{\beta}_1)$$

Hypothesis Testing

- Standard errors can also be used to perform hypothesis tests on the coefficients. The most common hypothesis test involves testing the null hypothesis of
$$H_0\text{: There is no relationship between } X \text{ and } Y$$
Versus the alternative hypothesis
$$H_A\text{: There is some relationship between } X \text{ and } Y$$
- Which mathematically corresponds to testing $H_0\colon \beta_1 = 0$ versus $H_A\colon \beta_1 \neq 0$ since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$ and X is not associated with Y
- T-statistic Test
    - If $SE(\hat{\beta}_1)$ is low, then small $\hat{\beta}_1$ values will present evidence to reject the null hypothesis. Also, if $SE(\hat{\beta}_1)$ is high, then only large $\hat{\beta}_1$ values will present evident to reject the null hypothesis.
    - To test the null hypothesis, we compute a t-statistic given by
$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$
Which measures the number of standard deviations that $\hat{\beta}_1$ is away from zero
    - This will have a t-distribtuion with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. This probability is called the p**-value**
- If the null hypothesis *cannot* be rejected, then the confidence interval for $\hat{\beta}_1$ *will* contain zero; if the null hypothesis *can* be rejected, then the confidence interval *will no*t contain zero

Assessing the Accuracy of the Model

- The **residual standard error** is an estimate of the standard deviation of ϵ. It is the average amount that the response will deviate from the true regression line. It is computed using the formula

$$RSE = \sqrt{\frac{RSS}{n-p-1}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  RSE is a good measure of lack of fit; if it is small, then predictions are close to true outcome values, if it is large, then predictions deviate from true outcome values

- R**-squared**
  - R-squared, or fraction of variance explained is
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$
    Where $TSS = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$ is the **total sum of squares.**
  - TSS measures the variance in response to Y, RSS measures the unexplained variability; hence $TSS - RSS$ measures the amount of variability in the response that is explained using the regression, and R^2-squared measures the proportion of variability in Y that can be explained using X.
  - It can be shown that in the *simple* linear regression setting that $R^2 = r^2$ where r is the correlation between X and Y

Multiple Linear Regression

- Here our model is
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$
- We interpret β_j as the *average* effect on Y of a one unit increase in X_j *holding all other predictors* fixed
- Interpreting Regression Coefficients
  - The ideal scenario is when the predictors are uncorrelated—a **balanced design**
    - Each coefficient can be estimated and tested separately
    - Interpretations such as "*a unit change in X_j is associated with a β_j change in Y, while all the other variables stay fixed*" are possible
  - Correlations amongst predictors cause problems
    - The variance of all coefficients tends to increase, sometimes dramatically
    - Interpretations become hazardous—when X_j changes, everything else changes
  - **Claims of causality** should be avoided for observational data
- Estimation and Prediction for Multiple Regression
  - Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

    - Like simple linear regression, the values that minimize RSS are the regression coefficient estimates

Important Regression Questions

- Is at least one predictor useful?
    - F-Statistic Test
        - We can use the F**-statistic**

$$F = \frac{\frac{TSS - RSS}{p}}{\frac{RSS}{n-p-1}} \sim F_{p,n-p-1}$$

          To test a null hypothesis that $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ versus the alternative that at least one β_j is non-zero

        - If the linear model assumptions are correct, $E\left\{\frac{RSS}{n-p-1}\right\} = \sigma^2$ and when H_0 is true, $E\left\{\frac{TSS-RSS}{p}\right\} = \sigma^2$. Thus, the F statistic will be 1 when H_0 is true. When H_a is true, then $E\left\{\frac{TSS-RSS}{p}\right\} > \sigma^2$ so the F statistic will be greater than 1.
        - When n (the number of training data points) is large, a F-statistic slightly larger than 1 may be enough to reject the null; if n is small, then a large F-statistic may be enough to reject the null
- Which variables are important?
    - The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets (of the set of predictors) and then choose between them based on some criterion that balances training error with model size
    - However, we often cannot examine all models since there are 2^p subsets a predictor set of cardinality p. We thus use two automated approaches:
    - Forward selection
        - Begin with the **null model**—a model that contains an intercept but no predictors
        - Fit p simple linear regressions and add to the null model the variable that results in the lowest RSS
        - Add to that model the variable that results in the lowest RSS amongst all two-variable models
        - Continue until some stopping rule is satisfied, for example, when all remaining variables have a p-value above some threshold
    - Backward selection
        - Start with all variables in the model
        - Remove the variable with the largest p-value, that is, the variable that is the least statistically significant

        - The new $(p-1)$ variable model is fit, and the variable with the largest p-value is removed
        - Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value defined by some significance threshold
    - Mixed selection
        - State with no variables in the model
        - Add the variables that provide the best fit (as in forward selection)
        - If at any point, the p-values for one of the variables in the model rises above a certain threshold, then we remove that variable from the model
        - Continue until all variables have a sufficiently low p-value and all variables outside the model would have a large p-value if added to the model
- How are our predictions?
    - There are three kinds of uncertainty associated with the prediction:
        - We have inaccuracy due to reducible error. We can compute a confidence interval for $\hat{Y}$ to determine how close our model is to the true model
        - We have **model bias** from assuming a linear model
        - We have irreducible error. We can compute a **prediction interval** for Y at a certain value of X to determine how $\hat{Y}$ will vary from Y. Such intervals are wider than confidence intervals because the incorporate the reducible and irreducible error. We interpret these intervals as saying that a certain percentage of intervals of this form will contain the true value of Y for a certain value of X

Qualitative Predictors

- Variables that are qualitative are called **categorical predictors** or **factor variables**
- Make dummy variables to represent being in a category or not

$$x_i = \begin{cases} 1\ if\ in\ category \\ 0\ if\ not\ in\ category \end{cases}$$

  So the resulting model will look like

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i\ if\ in\ category \\ \beta_0 + \epsilon_i\ if\ not\ in\ cateogry \end{cases}$$

- Interpreting Coefficients
    - β_0 is the value if not in the category
    - β_1 is the change in the value with respect to β_0 if the value is in the category
    - We can also change the 1,0 binary into $1, -1$ binary which will change the interpretation of the coefficients so that β_0 is a now a kind of average value, and β_1 is the change in value with respect to β_0 (positive change for in the category, a negative change for not in the category)

  - With more than two levels, we create $k - 1$ dummy variables for k categories. In the case that all dummy variables are zero, the variable takes on the **baseline** which is encompassed in β_0

Removing the Additive Assumption

- Instead of adding predictors, we can multiply them to represent a **synergy** or **interaction effect** which will be treated as its own predictor with its own β value. Such a model will look something like
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$$
Where β_3 is the coefficient of the **interaction term**
- Hierarchy Principle
  - Sometimes it is the case that an interaction term (the non-additive term) has a very small p-value, but the associated main effects (the coefficients of the single variables) do not
  - This case invokes is the **hierarchy principle**: if we include an interaction in a model, we should also include the main effects, even if the p-values associated with their coefficients are not significant
  - The rationale for this principle is that interactions are hard to interpret without main effects
  - Specifically, the interaction terms also contain main effects, if the model has no main effect terms

Non-linear Relationships

- **Polynomial regression** can be used to extend the linear model to accommodate non-linear relationships
- We can square, cube, etc. some of the predictors to make a “new” predictor that allows for different fits whilst still maintaining the linearity of the model

Potential Regression Problems

1. Non-linearity of the Data
   - Linear regression assumes there is a linear relationship between predictors and response
   - A residual plot is a good indicator of whether this is a fair assumption
   - If this assumption is an issue, it can be compensated for by using non-linear transformations of the predictors
2. Correlation of Error Terms
   - Correlation among the error terms may cause estimated standard errors to underestimate true standard errors
   - This can make confidence/ prediction intervals will be narrower than they should
3. Non-Constant Variance of Error Terms
   - Non-constant variances in the errors, or **heteroscedasticity**, that increases with an increase in the response can be an issue

- This can be compensated by transforming the response using a concave function like $\sqrt{Y}$ and $\log Y$
- We can also fit models by **weighted least squares** where weights are proportional to the inverse of variances

4. Outliers
   - Outliers can affect some values related to the regression model like RSE values
5. High Leverage Points
   - **High leverage points** are data points with outliers in the predictors that can drastically effect a model's coefficients
   - The **leverage statistic** quantifies an observation's leverage

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_{i'} - \bar{x})^2}$$

6. Collinearity
   - **Collinearity** refers to a situation in which two or more predictors are closely related to each other
   - **Multicollinearity** refers to a situation in which three or more predictors together are closely related, but pairwise none are
   - The variance inflation factor is a better way to assess collinearity. It is calculated according to

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Typically, a VIF close to 1 indicates little collinearity; a value between 5 and 10 indicates a problematic amount of collinearity

- Issues can be solved by dropping problematic variables that do not have too much of an undesirable impact on R^2

K-Nearest Neighbors Regression

- Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, represented by $\mathcal{N}_0$. It then estimates $f(x_0)$ using the average of all the training responses in $\mathcal{N}_0$. Stated mathematically,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

- The optimal K value will depend on the bias-variance tradeoff
- The **curse of dimensionality** describes when there are observations spread over so many dimensions that the nearest neighbor is actually very far away
- Typically, parametric methods will outperform non-parametric methods when there is a small number of observations per predictor

Chapter 4 – Classification

Classification

- Qualitative variables in an unordered set $\mathcal{C}$
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$
- Often we are more interested in estimating the probabilities that X belongs to each category in $\mathcal{C}$
- Can we use linear regression?
  - In the case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to **linear discriminant analysis** which we discuss later
  - Since in the population $E(Y|X = x) = \Pr(Y = 1|X = x)$ we might think that regression is perfect for this task
  - However linear regression might produce probabilities less than zero or bigger than one. **Logistic regression** is more appropriate.
  - Also, linear regression requires an ordering of outcomes and can imply that the difference between one outcome is the same as the difference between another, when in reality, such ordering and differences are inaccurate characterizations of the data
  - Summarizing, regression will not work because a) it cannot accommodate a qualitative response variable with more than two classes and b) it may provided meaningless estimates of $\Pr(Y|X)$

Logistic Regression

- In logistic regression (with two classes), we estimate $\Pr(Y = 1|X) = p(x)$ according to
$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$
- We use **maximum likelihood** to fit this model
- After some manipulation, we find that
$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}$$
Where the left-hand side is called the **odds**
- Taking the logarithm of both sides gives
$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$
called the **log odds** or **logit**.
- Unlike regression, a unit increase in X will cause the logit to increase by β_1 and not $p(X)$

Maximum Likelihood

- This is a method of fitting logistic regression models

- We seek estimates for β_0 and β_1 such that the predicted probability $\hat{p}(x_i)$ of default for each individual corresponds as closely as possible to the actual status of the response
- Mathematically, this is stated using the likelihood function

  $$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} \left(1 - p(x_{i'})\right)$$

  which we try to maximize
- Using this fitting, the z-statistic serve the same role as the t-statistic in the linear regression output; it tests the null hypothesis that $\beta_1 = 0$. The statistic is $z = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}$

Multiple Logistic Regression

- **Multiple logistic regression** concerns when there is more than one predictor
- The function for predicting probability has the symmetric form

  $$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- **Confounding** describes when there is correlation among the predictors; this can prove a niche point to navigate when interpreting results
  - Example
    - Consider data that tries to predict the probability of defaulting of a group of people based on their credit card balance, income, and student status
    - Over constant balances/ incomes, it may be that students default less than non-students
    - Thus, a β coefficient for student status (where student status is 1 if student) in a multiple logistic regression model that considers balance and income will be negative
    - However, because students overall have more balance than non-students, and high balance is associated with high probability for default, in a single predictor model, there will be a positive β coefficient for student status

Multinomial Regression

- Logistic regression with two classes is easily generalized to more than two classes after we select a single class to serve as the **baseline** (in this case, we choose class $Y = K$ as our baseline for no particular reason
- We replace the model according to

  $$\Pr(Y = k | X = x) = P(X = x) = \frac{e^{\beta_{0k} + \beta_{1k} X_1 + \cdots + \beta_{pk} X_p}}{\sum_{\ell=1}^{K} e^{\beta_{0\ell} + \beta_{1\ell} X_1 + \cdots + \beta_{p\ell} X_p}}$$

  For $k \in \{1, \ldots, K - 1\}$ and

  $$\Pr(Y = K | X = x) = P(X = x) = \frac{1}{1 + \sum_{\ell=1}^{K} e^{\beta_{0\ell} + \beta_{1\ell} X_1 + \cdots + \beta_{p\ell} X_p}}$$

- So for $k \in \{1, \ldots, K - 1\}$, we have

$$\log\left(\frac{\Pr(Y=k \mid X=x)}{\Pr(Y=K \mid X=x)}\right)=\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p$$

- Here there is a linear function for each of the K classes is also referred to as **multinomial regression**
- The log odds between any pair of classes will be the same regardless of which category is our baseline case, but the interpretation of our coefficients will change

Case-control sampling and logistic regression

- With case-control samples, we can estimate the regression parameters β_j accurately (if our model is correct); the constant term β_0 is incorrect
- We try to estimate whether some case will be true or not, for example, a disease
  However, there will be a large difference between π, the population prevalence and $\tilde{\pi}$, the prevalence of our sample. There is a large difference because for this testing, we get as many positive cases as we can find to compare to control groups. We often seek out more cases for our sample than proportionally exists in the population, especially for a rare disease
- We can correct the estimated intercept by a simple transformation

$$\hat{\beta}_0^*=\hat{\beta}_0+\log\frac{\pi}{1-\pi}-\log\frac{\tilde{\pi}}{1-\tilde{\pi}}$$

- Case control sampling is a faster and cheaper alternative to sampling from a large population and following them for a long period of time to see if they develop the condition or not. Case control sampling looks at people who already have the condition and those who do not

Generative Models

- Approach
  - Here the approach is to model the distribution of X in each of the classes separately and then use Bayes Theorem to obtain $\Pr(Y=k|X=x)$
  - When we use normal distributions for each class, this leads to linear or quadratic discriminant analysis
  - However, this approach is quite general, and other distributions can be used as well
- Why another approach?
  - When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. These methods do not suffer from this problem
  - If n is small and the distribution of the predictors X is approximately normal in each of the classes, approaches might be more stable/ accurate than the logistic regression model
  - These methods can naturally expand to more than two response classes, because it also provides low-dimensional views of the data
- **Bayes Theorem for Classification**
  - We have the classification version of Bayes Theorem is

$$p_k(x)=\Pr(Y=k|X=x)=\frac{\pi_k f_k(x)}{\sum_{l=1}^{K}\pi_l f_l(x)}$$

- Where $f_k(x) = \Pr(X = x|Y = k)$ is the **density** for X in class k. Here, we will use normal densities for these separately in each class
- $\pi_k = \Pr(Y = k)$ is the marginal or **prior** probability for class k; it is usually estimated as the fraction of training observations that belong to the k class
- We classify a new point according to which $p_k(x)$ is the highest
- The trouble is estimating $f_k(x)$
  - We use LDA, QDA, KNN neighbors, and naive Bayes to do so

Linear Discriminant Analysis (LDA) when $p = 1$

- We assume that $f_k(x)$ is normal; thus

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \mathrm{e}^{\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^2\right)}$$

  Where μ_k and σ_k^2 are the mean and variance parameters for the kth class

- We assume constant variance (σ^2) across all K classes. Thus, we have

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{\left(-\frac{1}{2\sigma^2}(x-\mu_k)^2\right)}}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{\left(-\frac{1}{2\sigma^2}(x-\mu_l)^2\right)}}$$

- The Bayes classifier involves assigning an observation $X = x$ to the class for which $p_k(x)$ is largest; this is equivalent to assigning the observation to the class where

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log{(\pi_k)}$$

  Is largest

- The Bayes decision boundary between class 1 and 2 would be the point for which $\delta_1(x) = \delta_2(x)$ which amounts to

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}$$

- In practice, the estimate parameters μ_k and π_k are not known, so **linear discriminant analysis** plugs in estimates for these parameters using

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

  Where n is the total number of training observations and n_k is the number of training observations in the kth class

- Evidently, the estimate for μ_k is the average of all the training observations from the kth class while σ is a weighted average of sample variances from each of the k classes
- When LDA is unable to make these calculations because of a lack of information, π_k is simple the proportion of training observations that belong to the kth class. In other words, $\hat{\pi}_k = n_k/n$

- LDA assigns x to the class k which maximizes the equation
$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$
- In summary, the LDA assumes observations within each class are normal with a class-specific mean, common variance, and puts these estimates into the Bayes classifier

Linear Discriminant Analysis (LDA) When $p > 1$

- We assume predictors are drawn from a multivariate normal distribution with a class specific mean vector (list of means for each predictor in each class) and a common covariance matrix
- Each individual predictor follows a one-dimensional normal distribution with correlation between each pair of predictors
- To indicate a p −dimensional random variable X with a multivariate normal distribution, we write $X \sim N(\mu, \mathbf{\Sigma})$ Here $\mu = E(X)$, the mean of vector X and $Cov(X) = \mathbf{\Sigma}$ is the $p \times p$ covariance matrix of X
- The multivariate normal density is defined as
$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)}$$
- The LDA classifier assumes observation from the kth class are drawn from a multivariate distribution $N(\mu_k, \mathbf{\Sigma})$ where μ_k is the class specific mean vector and $\mathbf{\Sigma}$ is the common covariance matrix
- The LDA classifier assigns X to the class k for which
$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log\pi_k$$
Is the largest. Despite this complex form, $\delta_k(x)$ is still linear and can be written in the form
$$\delta_k(x) = c_{k0} + c_{k1}x_1 + c_{k2}x_2 + \cdots + c_{kp}x_p.$$
- The Bayes decision boundary for class k and ℓ such that $k \neq \ell$ are the values x for which $\delta_k(x) = \delta_\ell(x)$ which is further expanded as
$$x^T \mathbf{\Sigma}^{-1} \mu_k - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k = x^T \mathbf{\Sigma}^{-1} \mu_l - \frac{1}{2} \mu_l^T \mathbf{\Sigma}^{-1} \mu_l$$
- When there are K classes, linear discriminant analysis can be viewed exactly in a $K - 1$ dimensional plot
    - This is because it essentially classifies to the closest centroid, and they span a $K - 1$ dimensional plane

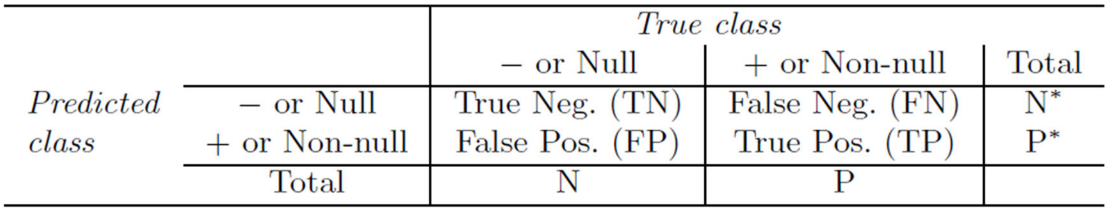
From $\delta_k(x)$ to Probabilities

- Once we have estimates $\hat{d}_k(x)$, we can turn these into estimates for class probabilities according to
$$\widehat{\Pr}(Y = k \mid X = x) = \frac{e^{\hat{\delta}_k(x)}}{\sum_{l=1}^{K} e^{\hat{\delta}_l(x)}}$$

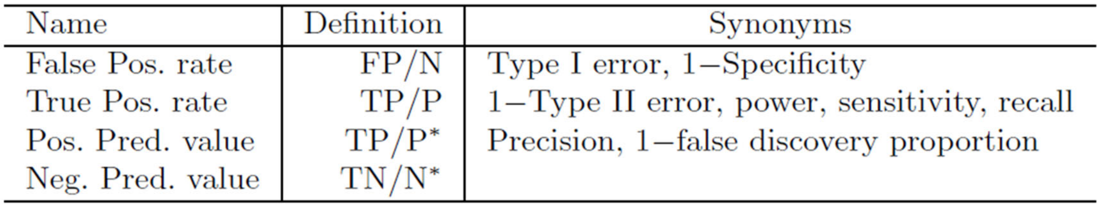
- Thus, classifying to the largest $\hat{\delta}_k(x)$ amounts to classifying to the class for which $\widehat{Pr}(Y = k|X = x)$ is the largest

Assessing Errors

- Metrics
  - **Sensitivity (true positive rate)**: the percentage of true positives that are classified as positive
  - **Specificity (true negative rate)**: the percentage of true negatives that are classified as negative (not positive)
  - **False positive rate**: The percentage of true negatives that are classified as positive (1-Specificity)
  - **False negative rate**: The percentage of true positives that are classified as negative (1-Sensitivity)
- A confusion matrix allows easy calculation of these four values
- The LDA has such low sensitivity since it will always have the smallest possible total number of misclassified observations. This can be compensated for by assigning observation x to class k only when $p(x)_k > \alpha$ for some threshold $\alpha \in [0,1]$ that corresponds to the needs of the model which are understood by **domain knowledge**
- To reduce the false positive rate, we increase the threshold. To reduce the false negative rate, we lower the threshold
- A **receiver operating characteristics (ROC) curve** plots sensitivity (true positive rate) vs false positive rate
- The **area under the curve (AUC)** gives the overall performance of a model as a number in $[0,1]$; the closer to 1, the better
- Some tables to help:



		True class		
		− or Null	+ or Non-null	Total
Predicted class	− or Null	True Neg. (TN)	False Neg. (FN)	N*
	+ or Non-null	False Pos. (FP)	True Pos. (TP)	P*
	Total	N	P	



Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

Quadratic Discriminant Analysis (QDA)

- QDA has the same goals as LDA, but with different assumptions
- QDA still assumes observations are drawn from a normal distribution and puts parameter estimates into Bayes' theorem, but, instead of assuming a constant covariance matrix like LDA, QDA assumes each class has its own covariance matrix; mathematically this means an observation from the k class is of the form $X \sim N(\mu_k, \mathbf{\Sigma}_k)$ where $\mathbf{\Sigma}_k$ is the covariance matrix for the kth class
- This means and observation with predictors x is assigned to the class for which
$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\Sigma_k| + \log\ \pi_k \\ &= -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log|\Sigma_k| + \log\ \pi_k\end{aligned}$$
Is the highest
- This function eventually appears as a quadratic function
- LDA vs QDA
    - Choosing LDA over QDA is another manifestation of the bias-variance tradeoff
    - LDA requires a total of $\frac{p(p+1)}{2}$ predicted parameters whereas QDA requires a total of $\frac{Kp(p+1)}{2}$ predicted parameters
    - LDA is much less flexible than QDA so it has a high variance, but can suffer from high bias if its strong assumptions are not met

Naïve Bayes

- Like LDA and QDA we are still attempting to estimate $f_k(x)$, the p-dimensional density function for an observation in the kth class
- For Naïve Bayes (NB), the assumption is made that within the k class, the p predictors are independent; mathematically, this means
$$f_k(x) = f_{k1}(x_1) \times f_{k2}(x_2) \times \ldots \times f_{kp}(x_p)$$
Where f_{kj} is the density function of the jth predictor among observations in the kth class
- Why the assumption?
    - Estimating a p dimensional density function is difficult because we must consider the **marginal distribution** of each predictor (the distribution of the predictors themselves) and the **joint distribution** (how the predictors are associated, summarized by a covariance matrix)
    - The covariances are hard to characterize and estimate
    - Thus, by assuming predictor independence within each class, we do not need to worry about the association estimates
    - This assumption is particularly helpful where n is not large enough relative to p for us to effectively estimate predictor distributions within each class

- Once this assumption is met, we put the assumption into Bayes Theorem for Classification which yields

  $$\Pr\,(Y = k \mid X = x) = \frac{\pi_k \times f_{k1}(x_1) \times f_{k2}(x_2) \times \cdots \times f_{kp}(x_p)}{\sum_{l=1}^{K} \pi_l \times f_{l1}(x_1) \times f_{l2}(x_2) \times \cdots \times f_{lp}(x_p)}$$

  For $k \in \{1,2,\dots,K\}$
- Estimating f_{kj} Using Training Data
  - For quantitative predictors X_j, we can assume that $X_j \mid Y = k \sim N(\mu_{jk}, \sigma_{jk}^2)$
    - This means that within each class, the jth predictor is drawn from a normal distribution
  - For quantitative predictors X_j, we can make a histogram (or use a smooth version of a histogram known as the **kernel density estimator**) for the observations of the jth predictor within each class and estimate $f_{kj}(x_j)$ as the fraction of the training observations in the kth class that belong to the same histogram bin as x_j
  - For a qualitative X_j, we can simple count the proportion of the training observations for the jth predictor corresponding to each class

Comparing Methods Analytically

- Comparing the methods analytically requires considering each approach in a setting with K classes
- As review, all methods work by assigning x to class k based on maximizing $\Pr\,(Y = k|X = x)$. Equivalently, we set K as the baseline class and assign an observation to the class that maximizes $\log\left(\frac{\Pr\,(Y=k\,|X=x)}{\Pr\,(Y=K|X=x)}\right)$ for $k = 1, \dots, K$
- Log Ratio Across Different Classes
  - For LDA, we see

    $$\log\left(\frac{\Pr(Y = k\,|X = x)}{\Pr(Y = K|X = x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j$$

    Where $a_k = \log\left(\frac{\pi_k}{\pi_K}\right) - \frac{1}{2}(\mu_k + \mu_K)^T\Sigma^{-1}(\mu_k - \mu_K)$ and b_{kj} is the jth component of $\Sigma^{-1}(\mu_k - \mu_K)$
  - For QDA, we see

    $$\log\left(\frac{\Pr\,(Y = k \mid X = x)}{\Pr\,(Y = K \mid X = x)}\right) = a_k + \sum_{j=1}^{p} b_{kj}x_j + \sum_{j=1}^{p}\sum_{l=1}^{p} c_{kjl}x_jx_l$$

    where a_k, b_{kj}, and c_{kjl} are functions of $\pi_k, \pi_K, \mu_k, \mu_K, \mathbf{\Sigma}_k$ and $\sum_K$
  - In the Bayes setting,

    $$\log\left(\frac{\Pr(\,Y = k \mid X = x\,)}{\Pr(\,Y = K \mid X = x\,)}\right) = a_k + \sum_{j=1}^{p} g_{kj}(x_j)$$

Where $a_k = \log\left(\frac{\pi_k}{\pi_K}\right)$ and $g_{kj}(x_j) = \log\left(\frac{f_{kj}(x_j)}{f_{Kj}(x_j)}\right)$

- What This Tells Us
    - LDA is a special case of QDA with $c_{kjl} = 0$ for all $j = 1, \dots, p, l = 1, \dots, p,$ and $k = 1, \dots, K$
    - Any classifier with a linear decision boundary is a special case of naïve Bayes with $g_{kj}(x_j) = b_{kj}x_j$. This means LDA is a special case of naïve Bayes; LDA assumes the features have a common within-class covariance matrix while naïve Bayes instead assumes independence of the features
    - If we mode $f_{kj}(x_j)$ in the naïve Bayes classifier using a one-dimensional normal distribution then we end up with $g_{kj}(x_j) = b_{kj}x_j$ where $b_{kj} = (\mu_{kj} - \mu_{Kj})/\sigma_j^2$. In this case, naïve Bayes is a special case of LDA where $\mathbf{\Sigma}$ is restriced to be a diagonal matrix with jth diagonal element equal to σ_j^2
    - Neither QDA nor naïve Bayes is a special case of the other. Naïve Bayes produces a more flexible fit, but can only make additive fits. Instead, QDA included multiplicative terms, giving it the potential to be more accurate in settings where interactions among predictors are important in classification
    - Some methods will be better than others in certain scenarios keeping in mind the bias-variance trade-off
- How Logistic Regression Relates
    - For logistic regression, we see

      $$\log\left(\frac{\Pr\,(Y = k \mid X = x)}{\Pr\,(Y = K \mid X = x)}\right) = \beta_{k0} + \sum_{j=1}^{p} \beta_{kj}x_j$$

      Which shows that the log ratio, like LDA, is a linear function of the predictors
    - The coefficients in logistic regression are estimates for $\pi_k, \pi_K, \mu_k, \mu_K$ and $\mathbf{\Sigma}$ obtained by assuming predictors follow a normal distribution within each class
    - Logistic regression chooses coefficients to maximize the likelihood function
    - Thus, logistic regression outperforms LDA when the assumptions of LDA do not hold, but LDA does better when its assumptions do hold
- KNN
    - KNN will dominate over LDA and logistic regression when the decision boundary is highly non-linear, n is very large, and p is small
    - In settings with a modest n or a non-small p, QDA may be preferred to KNN since it can still provide non-linear decision boundaries
    - KNN does not tell us which predictors are important

Generalized Linear Models

- For data that is either non-quantitative or qualitative, linear regression and classification methods have some unignorable errors
- A common method for modeling these types of scenarios is the Poisson regression
- Poisson Distribution
  - Suppose a random variable Y takes on nonnegative integer values. If Y follows a Poisson distribution, then
  $$\Pr(Y = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0,1,2,\dots$$
  Where $\lambda = E(Y) = Var(Y)$
  - This distribution is typically used to model counts
- Poisson regression models the mean λ as a function of predictors which is made into a model by
$$\lambda(X_1, \dots, X_p) = e^{\beta_0+\beta_1 X_1+\dots+\beta_p X_p}$$
Or equivalently
$$\log\left(\lambda(X_1, \dots, X_p)\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$
where β_i are the parameters to be estimated
- These parameters are estimated according to the likelihood approach from logistic regression using the likelihood function
$$\ell(\beta_0, \beta_1, \dots, \beta_p) = \prod_{i=1}^{n} \frac{e^{-\lambda(x_i)}\lambda(x_i)^{y_i}}{y_i!}$$
- Distinctions Between Linear and Poisson Regression
  - Interpretation – unlike in linear, in the Poisson regression, we must interpret unit increases in predictors as affecting the response by a factor of e^{β_i} not just β_i
  - Mean-variance relationship - Unlike in linear regression, the assumption of a Poisson distribution, the variance changes with the mean (since $E(X) = Var(X)$) whereas in linear regression, the variance is assumed constant
  - Non-negative fitted values – There are no negative predictions using the Poisson regression model
- Generalized Linear Models Generalized
  - Each approach uses predictors to predict a response Y as a function of given predictors by modeling the expected value of Y (mean) as a function of predictors
  - The function that applies a transformation to $\mu = \mathrm{E}(Y \mid X_1, \dots, X_p)$ so that the transformed mean is a linear function of the predictors is known as the **link function** η
  - The link functions
    - For linear regression: $\eta(\mu) = \mu$
    - For logistic regression: $\eta(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
    - For Poisson regression: $\eta(\mu) = \log(\mu)$
  - The normal distribution, Bernoulli, and Poisson distributions are from a wider class of distributions known as the **exponential family**

- In general, we can preform regression by modeling Y as coming from a particular member of the exponential family and transforming the mean of the response so that the transformed mean is modeled as $\eta\left(\mathrm{E}\left(Y \mid X_1, \dots, X_p\right)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
- A regression model that follows this general recipe is known as a **generalized linear model**

Chapter 5 – Resampling Methods

Resampling Methods

- Resampling methods refit a model of interest to sample formed from the training set, in order to obtain additional information about the fitted model
- They can provide estimates of test-set prediction error and the standard deviation and bias of our parameter estimates
- Two common methods are cross-validation and the bootstrap
    - Cross validation is used to estimate test error with different statistical learning methods and to selected appropriate parameter values
    - The bootstrap is helpful for providing a measure of accuracy of a parameter estimate
- Methods are used for **model assessment**, which evaluates how a model performs, while others are used for **model selection**, which evaluates the appropriate level of flexibility a model should have

Cross Validation

- Training Error vs Test Error
    - The **test error** is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method
    - In contrast, the **training error** can be easily calculated by applying the statistical learning method to the observations used in its training
    - But the training error rate often is quite different from the test error rate, and in particular the former can dramatically underestimate the latter
    - Best solution for estimating test error is a large designated test set, but they are often not available, and the quantity that is desired to be predicted is not always known
    - There is a class of methods that estimate the test error by **holding out** a subset of the training observations from the fitting process, and then applying the statistical learning methods to those held out observations

The Validation-set Approach

- Here we randomly divide the available set of samples into two parts of comparable size, the **training set** and a **validation** or **hold-out** set
- The model is fit on the training set and the fitted model is used to predict responses for the observations in the validation set
- The resulting validation-set error provides an estimate of the test error typically assessed using MSE in the case of a quantitative response and misclassification rate in the case of a qualitative response

- Drawbacks
    - The validation estimates can be highly variable since the data set is split randomly. Estimates resulting from the validation set can be highly dependent on which exact data is included in the validation set compared to the training set
    - In the validation approach, only a subset of actual observations, the training set, are used to fit the model. This suggests the validation set error may tend to overestimate the test error for the model fit on the entire data set since when less data is used to create a model, it typically preforms worse

Leave-One-Out Cross-Validation (LOOCV)

- Similar philosophy to validation set approach, but seeks to assuage some of its drawbacks
- Like validation set approach, data is split into two parts, except the validation part is a now a singular observation
- For a set of n observations, the model is trained on the training set of $n-1$ observations, and then tested on the singular observation the MSE is then computed on this one observation to estimate the test error
- This procedure is repeated n times, each time leaving out a different observation until all observations have been the validation observation. Then the n computed MSE estimates are averaged for a final estimate of the model's test error:
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$
- Advantages
    - The estimate from LOOCV has far less bias than the validation set approach because each observation is considered in the error estimate, in contrast to validation-set which considers only a subset of the observations
    - There is also low variance since LOOCV is not affected by the split between validation and training data—each observation will eventually be part of the validation data
- One disadvantages is that LOOCV can be computationally expensive since it requires fitting a model n times. If the model chosen is complicated, computing time can indeed grow quickly. However, with least-squares or polynomial regression, it can be proven that
$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$
where $\hat{y}_i$ is the ith fitted value from the original least-squares fit and h_i is leverage as defined in Chapter 3

k-Fold Cross Validation

- This is an alternative to LOOCV
- Now, we divide the data into k sets of roughly equal size. We select $k-1$ of these sets to use for training and then one of these sets to use as a validation set; the MSE is computed on this validation set

- Like LOOCV, this method is then repeated k times so that each of the k fold gets a turn as the validation set. The final test error is similarly estimated according to

$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

- Typically, $k = 5$ or $k = 10$, where $k = n$ is the special case of LOOCV. Since models only need to be made k times, the computational power required to perform such validation can be much lower than LOOCV while still maintaining some of the bias-variance tradeoff qualities
- **The Bias-Variance Tradeoff**
    - LOOCV will have less bias than k-fold cross validation since the training set LOOCV uses to fit the model each time is about the same size as the entire data set
    - However, LOOCV does not "shake up" the data as much, and hence, each model is fit on almost identical data, meaning the estimated test error from each model is highly correlated. Taking the mean of highly correlated data has a higher variance than taking the mean of less-correlated data, so the test error estimate from LOOCV has higher variance than does k-fold cross validation
    - Hence, to balance this bias-variance tradeoff $k = 5$ or $k = 10$ has been shown to best balance bias and variance

Cross-Validation and Classification Problems

- For classification problems, cross validation works the same as it does for regression, except now, instead of MSE, the assessed metric is the number of misclassified observations. For LOOCV, the overall test error takes the form

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i$$

  where Err_i is the number of times $y_i \neq \hat{y_i}$. The k-fold cross validation and validation set error rates are defined analogously

Bootstrap

- The **bootstrap** is a flexible and powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator or statistical learning method
- It can provided an estimate of the standard error of model parameters
- Rather than repeatedly obtaining independent data sets from the population, which is sometimes impossible, we instead obtain distinct data sets by repeatedly sampling observations from the original data set with replacement
- Each of these bootstrap data sets is created by sampling with replacement and is the same size as our original dataset
- Procedure
    - Denoting the first bootstrap data set by Z^{*1}, we use Z^{*1} to produce a new bootstrap estimate for α which we call $\hat{\alpha}^{*1}$

  - This procedure is repeated B times for some large value B in order to produce B different bootstrap data sets $Z^{*1}, Z^{*2}, \dots, Z^{*B}$ and B corresponding α estimates $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \dots, \hat{\alpha}^{*B}$
  - We estimate the standard error of these bootstrap estimates using the formula

$$\mathrm{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B} \left(\hat{\alpha}^{*r} - \bar{\hat{\alpha}}^{*}\right)^2}$$

  - This serves as an estimate of the standard error of $\hat{\alpha}$ estimated from the original data set
- In more complex data situations, figuring out the appropriate way to generate bootstrap samples can require some thought, for example, when the data is correlated or not IID
- We can instead create blocks of consecutive observations, and sample those with replacements. Then we paste together sampled blocks to obtain a bootstrap dataset
- Can bootstrap estimate prediction error?
  - In cross validation, each of the K validation folds is distance from the other $K - 1$ folds used for training—there is no overlap
  - To estimate prediction error using the bootstrap, we could think about using each bootstrap dataset as our training sample, and the original sample as our validation sample
  - But each bootstrap sample has significant overlap with the original data. About two-thirds of the original data points appear in each bootstrap sample
  - This will cause bootstrap to seriously underestimate the true prediction error
  - Can partly fix this problem by only using predictions for those observation that did not (by chance) occur in the current bootstrap sample
  - But the method gets complicated, and in the end, the cross-validation provides a simpler, more attractive approach for estimating prediction errors

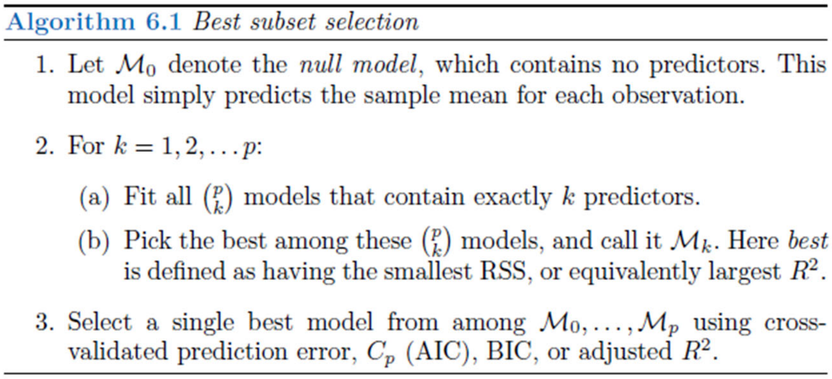
Chapter 6 – Linear Model Selection and Regularization

Linear Models Alternative to Least Squares

- While more flexible model types exists, generally, linear models offer inference and interpretability advantages while preforming comparatively well to non-linear methods
- Alternative fitting methods to least squared can yield better **prediction accuracy** and **model interpretability**
    - Prediction accuracy – Least squares will have low bias if the underlying relationship is approximately linear, and if the number of observations is vastly larger than the number of predictors, the least-squares estimates will have low variance. However, if this relationship does not hold, then there can be much variability and overfitting. If we constrain or shrink the estimated coefficients, then we can reduce variance at little cost to bias
    - Model interpretability – sometimes, irrelevant variables are included in models, leading to these models having an unnecessarily complex structure. If we can find these variables and remove them (set their coefficients to zero) using **feature selection** or **variable selection**, then our model will be more interpretable
- There are three alternative to least squares
    - Subset selection – we identify the subset of predictors believed to be related to the response, fitting a least-squares model to these predictors
    - Shrinkage – we fit a model with all predictors and then shrink their estimated coefficients to zero
    - Dimension reduction – This approach involves projecting the p predictors into an M-dimensional subspace where $M < p$. This is achieved with linear algebra techniques to compute M different linear combinations of variables which are then used as predictors to fit a linear regression model by least-squares

Best Subset Selection

- The Algorithm



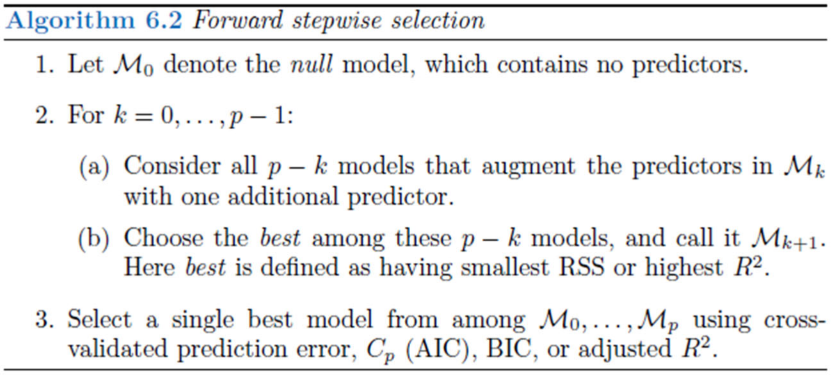
Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
    (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
    (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

  - Initially, we have 2^p models to consider, but because of step 2, we only have $p + 1$ possible models
  - We must select from these $p + 1$ models with care, as we are not only looking at RSS or R^2 on the training set, as this would produce and overfit model
  - In the case of logistic regression, instead of using RSS to select models in step 2, we use the deviance, which is negative two times the maximized log-likelihood; a smaller deviance indicates a better fit
- Evidently, there are computation limitations to this method since the number of possible models grows rapidly as p increases. For a very high p value, best subset selection cannot be applied, or the search space is so large that the found model may prove useless

Forward Subset Selection

- An alternative to best subset selection
- The Algorithm



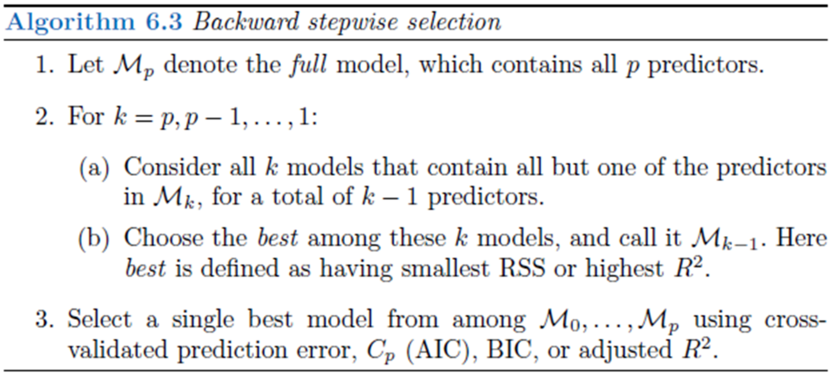
Algorithm 6.2 *Forward stepwise selection*

```
1. Let $\mathcal{M}_0$ denote the null model, which contains no predictors.
2. For $k = 0, \ldots, p - 1$:
    (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$
        with one additional predictor.
    (b) Choose the best among these $p - k$ models, and call it $\mathcal{M}_{k+1}$.
        Here best is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

  - Unlike best subset selection, forward stepwise selection involves fitting one null mode, along with $p - k$ models in the kth iteration for $k = 0, \ldots, p - 1$. This amounts to a total of $1 + \sum_{k=0}^{p-1} (p - k) = 1 + p(p + 1)/2$ models
- Though offering computational advantage, forward subset selection does not guarantee selection of the best model from all 2^p models
- Forward stepwise selection can be used when $n < p$, but it is only possible to construct the models up to the point when the index $i = p$ since for the models with $i > p$, the least squares solution will not yield a unique solution (because of reasons relating to linear algebra)

Backwards Stepwise Selection

- Another alternative to best subset selection
- The Algorithm



Algorithm 6.3 *Backward stepwise selection*

```
1. Let $\mathcal{M}_p$ denote the *full* model, which contains all $p$ predictors.
2. For $k = p, p-1, \ldots, 1$:
    (a) Consider all $k$ models that contain all but one of the predictors
        in $\mathcal{M}_k$, for a total of $k-1$ predictors.
    (b) Choose the *best* among these $k$ models, and call it $\mathcal{M}_{k-1}$. Here
        *best* is defined as having smallest RSS or highest $R^2$.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-
   validated prediction error, $C_p$ (AIC), BIC, or adjusted $R^2$.
```

  - Like forward stepwise selection, there are a total of $1 + \sum_{k=0}^{p-1} (p-k) = 1 + p(p+1)/2$ models searched
- Like forward stepwise selection, finding the best subset result is not guaranteed
- Backward stepwise selection will have issues preforming when $n < p$ since it starts with all p predictors and hence will not produce a unique least squares solution for the first model
- Hybrid approaches combine the computational efficiency of forward and backwards stepwise selection, adding relevant variables and dropping irrelevant ones, while still having available some of the vast possibilities of best subset selection

Choosing the Optimal Model

- We need to find some way to assess the best model from all the options given to us by subset, forward, and backward selection. We cannot use RSS or R^2 since this will contribute to overfitting, so we must estimate the test error by making adjustments to the training error to account for bias or using a validation set approach as from Chapter 5
- **Adjustments to Training Error**
  - There are methods we can use to adjust training error, which is typically an underestimate of test error, so it mimics testing error
  - C_p
    - This method estimates test MSE using the equation

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2)$$

      where $\hat{\sigma}^2$ is an estimate of the variance of error associated with each response (typically estimated using the full model with all predictors) and d is the number of predictors in each model
    - Essentially, this adds a penalty to the training RSS based on the number of predictors, compensating for the lower RSS, and thus more underestimate of testing error, given by more predictors
    - The value with lowest C_p is typically the one with lowest test error

  - Akaike information criterion (AIC)
    - This is defined for models fit by maximum likelihood, and when this is the same thing as least squared, we have AIC is the same as C_p
  - Bayesian information criterion (BIC)
    - Similar to two above methods, is given by

      $$BIC = \frac{1}{n}(RSS + log\ (n)d\hat{\sigma}^2)$$

      with same variable definitions
    - This places a heavier penalty than C_p on models with many variables
    - Smaller value also indicates lower test error
  - Adjusted R^2
    - For a least squares model with d variables, we have

      $$Adjusted\ R^2 = 1 - \frac{RSS\ /(n-d-1)}{TSS\ /(n-1)}$$

    - Here, a higher value indicates a better model
    - The reasoning behind adjusted R^2 is that, like normal R^2, it will increase with more variables being added, but only up until a certain point when noise variables added will not decrease RSS enough to make up for the requisite increase brought on by increasing d
- Validation and Cross Validation
  - Chapter 5 methods could be used to estimate test error
  - This makes fewer assumptions about the underlying structure of the model than training error adjustment options does
  - Validation and cross validation also require less information that might be hard to get, like the number of predictors or error variance
  - There may be a significant drop off in the decreased test error once a certain threshold of parameters are used in the model. To avoid including more unnecessary predictors for trivial drops in test error, we use the **one-standard-error** rule which selects the smallest model for which the estimated test error is within one standard error of the lowest test error on the curve.

Shrinkage Methods

- Such methods fit a model using all predictors, and then seek to constrain or regularize the coefficient estimates, forcing the coefficient estimates to zero
- Shrinking the coefficient estimates significantly reduce their variance; techniques of doing this are ridge regression and the lasso

Ridge Regression Shrinkage Method

- Very similar to least squares, except the coefficients are estimated by minimizing a slightly different quantity, namely the quantity
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \mathrm{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2$$
where $\lambda \geq 0$ is a tuning parameter
- The second term is called the **shrinkage penalty** and is small when the coefficients are close to zero and large when they are not; how much of a role it plays is entirely determined by the choice of λ
- Choosing the proper λ is important and should be done to minimize the test error
- It is best to apply ridge regression after standardizing predictors using the formula
$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}}$$
since the actual coefficient magnitudes matter significantly. This puts all coefficients on the same scale
- The advantage of ridge regression over least squares is another manifestation of the bias-variance tradeoff. With an increased λ, the flexibility of the model decreases, decreasing variance and increasing bias
- In general, where the relationship between the response and the predictors is close to linear, from least squares, there will be low bias but high variance, meaning a slight change in the training data may produce vastly different models. This is especially true when the number of variables is almost as large as the number of observations. Ridge regression can sacrifice small amounts of bias for large amounts of variance
- Since for any fixed value λ, ridge regression fits only a single model, there is much advantages in terms of computational power
- One disadvantage is a ridge regression model will always include all predictors; even ones with minimized coefficients can make interpretability hard

The Lasso Shrinkage Method

- Lasso is an alternative to ridge regression that allows the creation of models without all variables
- The lasso coefficients minimize the quantity
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = \mathrm{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|$$
which is evidently similar to ridge regression except the penalty term now has $|\beta_j|$ instead of β_j^2.

- This shrinks coefficient estimates towards zero, like ridge regression, except now, coefficients can actually take values of zero, meaning the models are sparse (containing less predictors) and thus more interpretable
- In situations where most coefficients are not in fact zero, lasso performs worse since it assumes some coefficients are, yet in situations where some or most coefficients are zero, ridge regression performs worse because it always includes all predictors. Hence, in general, when there is a small amount of predictors with large coefficients, lasso performs better, yet where response is a function of many predictors, each with similar coefficients, ridge regression will perform better
- Cross validation can help determine which model is better for a data set

Selecting the Tuning Parameter

- Cross validation provides an excellent way to determine the best λ value. We give a grid of λ values to the model and compute the cross validation testing errors for each, selecting the value with the lowest error
- If the optimal fit shows a relatively small λ value, this means penalizing predictors is not that important, meaning least squares might fare better

Dimension Reduction Methods

- Dimension reduction methods transform the original predictors and then fit a least squares model using the transformed variables
- If we let $Z_1, Z_2, \dots, Z_M$ represent $M < p$ linear combinations of our original p predictors, which means $Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$, for some ϕ_{jm} constants, we can then fit the linear regression model $y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, i = 1, \dots, n$ using least squares
- If we choose the ϕ_{jm} constants wisely, then such dimension reduction can often outperform least squares regression
- The term dimension reduction comes form that this approach reduces the problem of estimating the $p + 1$ coefficients to the simpler problem of estimating the $M + 1$ coefficients
- Dimension reduction serves to constrain the estimated β_j coefficients, having a potential to bias the estimates, but where predictors are large relative to the number of observations, selecting a sufficiently low value of p can significantly reduce the variance of the coefficients
- Dimension reductions work in two steps; first, the transformed predictors $Z_1, Z_2, \dots, Z_M$ are obtained. Second, the model is fit using these M predictors. The choice of the values of $Z_1, Z_2, \dots, Z_M$ can be achieved in different ways, two of which are principal components and partial least squares

Principal Components Analysis (PCA)

- PCA seeks to reduce the dimension of an $n \times p$ data matrix. The first principal component direction of the data is that along which the observations vary the most
- Another interpretation is that the first principal component vector defines the line that is as close as possible to the data

- **Principal Components Regression (PCR)**
    - Principal components regression involves constructing the first M principal components and then using them as predictors in a linear regression model that is fit using least squares
    - The key idea here is that a small number of principal components suffices to explain most of the variability in the data and the relationship with the response, or in other words, that the directions in which predictors show the most variation are the directions associated with the response
    - If these assumptions hold, we will outperform least squares and by estimating only a few coefficients, we mitigate overfitting
    - As more principal components are used, the bias decreases but variance increases
    - This is not a feature selection method, as each principal component used in regression is a linear combination of all p of the original features
    - The number of principal components M is often chosen by cross validation
    - It is recommended to standardize each predictor before generating principal components to ensure that all variables are on the same scale
- A drawback from these methods is that they are unsupervised and do not consider the response; there is no guarantee that the directions to explain predictors will best explain responses

Partial Least Squares (PLS)

- Like PCR, PLS first identifies a new set of features $Z_1, Z_2, \ldots, Z_M$ that are linear combinations of the original features and then fits a linear model via least squares using these new features
- However, unlike PCR, PLS makes use of the response in order to identify new features—it finds directions that explain both predictors and response
- These directions are computed by first standardizing predictors, and setting the Z_j coefficients to the coefficients from simple linear regression of the response onto the predictor. Hence, highest weights are placed on variables more strongly related to the response
- To identify the second direction, we adjust each of the variables in the first direction by regressing each variable on the first direction and taking residuals. These residuals represent the information that has not been explained by the first direction, which goes into determining the second direction. This is repeated M times and then we use least squares to fit a linear model to predict the response
- The value of M is predicted using cross-validation

Chapter 7 – Moving Beyond Linearity

Non-Linear Models

- The assumption of model linearity is always an approximation, and can only be improved so far for linear models that do not have an underlying linear structure
- We can relax the linear assumption, allowing for more flexibility while still maintaining interpretability

Polynomial Regression

- Here, we replace the standard linear model with a polynomial function given by
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$$
- For a large enough degree, polynomial regression allows us to produce an extremely non-linear curve. Typically, we do not choose a degree greater than 3 or 4 since a greater degree could result in a model that is too flexible and this overfits
- Coefficients can still be estimated using least squares

Step Functions

- Instead of imposing a global structure on our model, as is done in linear and polynomial regression, we can use step functions. These break the predictor region into binds and fit a different constant to each bin, converting a continuous variable into an ordered categorical variable
- We create cutpoints $c_1, c_2, \dots, c_K$ in the range of X and then construct $K + 1$ new variables
$$\begin{aligned} C_0(X) &= I(X < c_1) \\ C_1(X) &= I(c_1 \leq X < c_2) \\ C_2(X) &= I(c_2 \leq X < c_3) \\ &\vdots \\ C_{K-1}(X) &= I(c_{K-1} \leq X < c_K) \end{aligned}$$
Where I is an **indicator function** returning 1 if a condition is met and 0 otherwise. These are sometimes called **dummy variables**.
- We then use least squares to fit a linear model using $C_1, C_2, \dots, C_K$ as predictors, yielding
$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \epsilon_i$$

Basis Functions

- A basis function approach creates a family of functions or transformations that can be applied to a variable X according to $b_1(X), b_2(X), \dots, b_K(X)$
- Instead of fitting a linear model in X, we fit the model
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \beta_3 b_3(x_i) + \cdots + \beta_K b_K(x_i) + \epsilon_i$$
- These basis function are fixed and known, meaning we can treat each function of a predictor as its own variable and use least squares to fit the model
- Actually, polynomial regression is a special case where the basis functions are raising the predictors to some power. Likewise, step functions are a special case where the basis functions are indicator functions

- Many alternatives are available for these functions, including wavelets and Fourier series

Regression Splines

- Piecewise polynomial regression is a type of regression spline that involves fitting separate low-degree polynomials over different regions of predictors instead of one high degree polynomial over all data
- Polynomial models are predicted as in polynomial regression, except now, the coefficients can changed depending on the region. For example, if we have a **knot** at c (a point where the graph changes) the model would look like
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c \end{cases}$$
- Using more knots leads to more flexible models. With K knots, we will have $K + 1$ different functions fit on the $K + 1$ different regions. We can make these functions of a low polynomial degree (even linear or constant) and still have a flexible model
- To avoid graphs that look ridiculous, we can impose certain restrictions on our graphs. For example, we can impose a restriction that the graphs must be continuous at the knots. We can impose a restriction forcing the derivatives of the functions to be constant so we have a smooth looking graph
- A cubic spline with the condition that it is twice continuously differentiable is called a **natural spline**
- The total number of degrees of freedom we have is the number of knots minus the number of constraints
- Choosing the Number and Locations of the Knots
  - The regression region is more flexible where there are many knots; hence, it might be best to place knots in the regions where the function might vary most rapidly and place fewer knots where it seems to be stable
  - We can use cross validation to decide where to place the knots
- Comparison to Polynomial Regression
  - Regression splines often give superior results to polynomial regression since they can produce flexible fits without very high powers
  - We can also choose regions where the graph should be more flexible and maintain regions where the graph is more stable

Smoothing Splines

- The goal here is to find a function that makes RSS small but is smooth to avoid being too flexible
- One method is to find some function g that minimizes

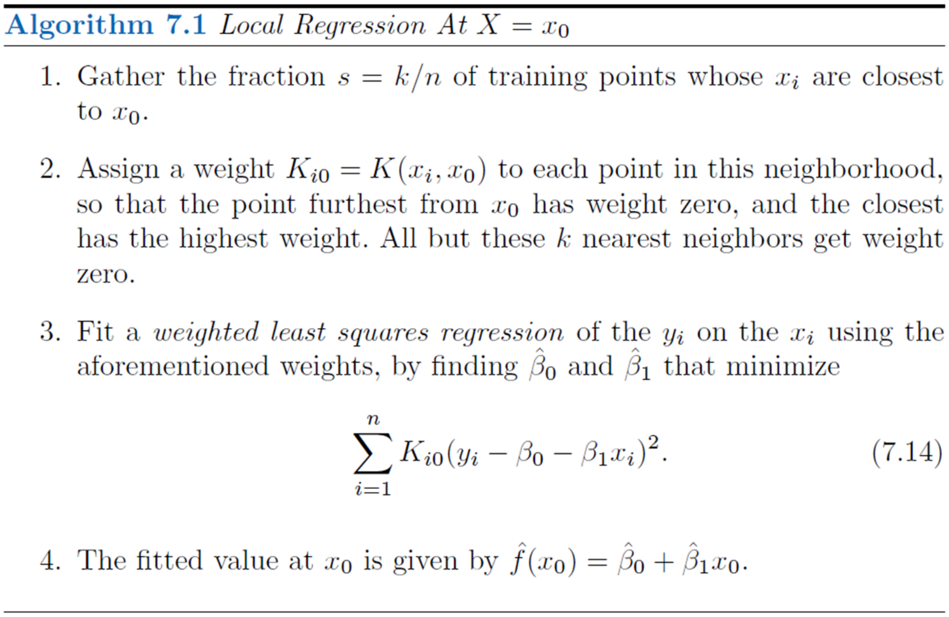
$$\sum_{i=1}^{n} \left(y_i - g(x_i)\right)^2 + \lambda \int g''(t)^2 dt$$

where λ is a nonnegative tuning parameter. The function g is known as a smoothing spline

- This function takes a $Loss + Penalty$ form like in ridge regression, lasso. The first term is the error that is meant to be minimized while the second penalized having a wildly changing second derivative. Hence, this penalty seeks to minimize the roughness of the function g and regulates the degrees of freedom of the function, affecting the bias-variance tradeoff

Local Regression

- This method involves computing the fit at a target point x_0 using only the nearby training observations
- The Algorithm



Algorithm 7.1 *Local Regression At* $X = x_0$

1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.
3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \qquad (7.14)$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

  - The choices that go into local regression are whether the local models be linear, polynomial, or some other type. Another choice is the weighting parameter K for each point.
  - Yet another choice is the span s, or the proportion of the total data points that are used to compute the local regression at a point. A small s will make the functions local and more wiggly, whereas a large value will lead to a more global fit
  - We can use cross validation to choose these values
- One useful generalization of such **varying coefficient models** as local regression is that established models can be adapted to new data
- Also, such methods can fit models that are global in some variables but local in others. It is also useful for models that are local in a pair of variables

Generalized Additive Models

- GAMs can be used to model responses from multiple predictors
- Such models provide a framework for extending a standard linear model by allowing non-linear functions of each of the variables while maintaining additivity. These models can be applied to both quantitative and qualitative responses

- We can use all the above methods for this chapter to generate a function $f_j(x_{ij})$ for each predictor and sum them together according to

$$\begin{aligned} y_i &= \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i \\ &= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i \end{aligned}$$

  to get a final model
- Pros
  - Allow us to fit a non-linear model to each predictor so we can model non-linear relationships that standard linear regression will miss
  - Non-linear fits can make more accurate predictions for the response
  - We can examine the effect of each predictor on the response while holding all other variables fixed since the model is additive
  - The smoothness of the function can be summarized via degrees of freedom
  - GAMs can also be used for classification models where, like logistic regression, they model the logarithm associated with the response
- Cons
  - The model is restricted to additive models which can miss many important interactions

Chapter 8 – Tree Based Methods

Tree Based Methods

- These methods involve stratifying or segmenting the predictor space into a number of simple regions
- To make predictions for these regions, we typically use the mean or mode of the response value of these regions
- The rules for splitting these regions can be summarized by decision trees
- Tree methods are simple and useful for interpretation, but are not competitive in terms of predication accuracy with the best supervised learning approaches

Regression Trees

- Regions in which the predictor space is split up are referred to as the **leaves** or **terminal nodes** of the tree
- The places where the regions split are referred to as **internal nodes**. Segments of the tree connecting these nodes are **branches**
- Process for Building
  - There are two main steps
  - First, we divide the predictor space into distinct, non-overlapping regions $R_1, R_2, \dots R_J$
  - Next, for every observation that falls into a particular region, we make the same prediction, whether that be the mean of the response values for that region
- Constructing regions
  - We choose to divide the regions into high-dimensional rectangles or boxes for simplicity and ease of interpretation
  - The goal is to find boxes $R_1, \dots, R_j$ that minimize the RSS given by

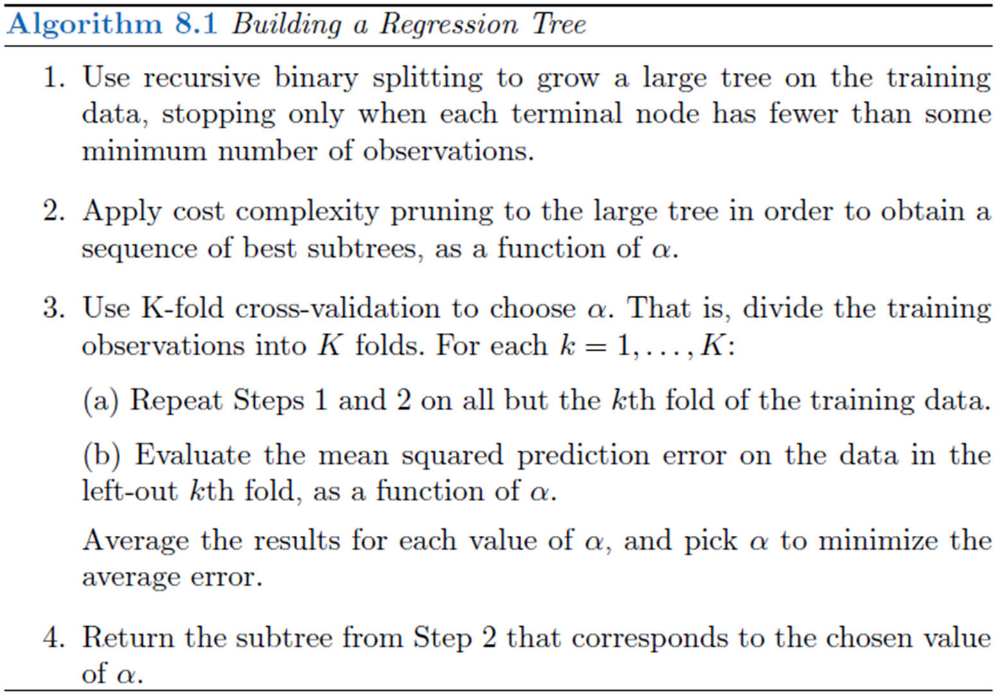
    $$\sum_{j=1}^{J} \sum_{i \in R_j} \left(y_i - \hat{y}_{R_j}\right)^2$$

    Where $\hat{y}_{R_j}$is the mean response for the training observations within that box
  - Since we cannot consider every possible region of the predictor space, we use a top-down or greedy approach where we begin at the top of the tree and the best split is made at each particular step. This method is known as **recursive binary splitting**
  - We first select the predictor X_j and the cut point s such that splitting the predictor space into the regions $\{X \mid X_j < s\}$ and $\{X \mid X_j \geq s\}$ that leads to the greatest possible reduction in RSS
  - We keep doing this process of splitting at each best predictor within the resulting regions
  - This process continues until a stopping criteria is reached, such as a minimum number of observations in each region, a specified number of regions, or a threshold RSS value

- Tree pruning
    - The process described above may overfit the data by creating too many regions, leading to poor test set performance. Smaller trees might have lower variance and better interpretation at the cost of some bias
    - By setting some conditions as desceibed above, we can ensure the tree does not become too complicated, but the drawback is this may prevent the tree from getting to a split that is actually meaningful
    - Hence, the best method is to build a vary large initial tree and prune it back to a reasonable subtree
    - We can estimate test error using cross validation to determine the best pruning methods. But considering every possible subtree would be too cumbersome
    - **Cost complexity pruning**, also known as **weakest link pruning**, gives us a good way to prune. Rather than considering all possible subtrees, we consider a sequence of trees indexed by some non-negative tuning parameter α. For each value of α we have a subtree T of our original tree such that

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} \left(y_i - \hat{y}_{R_m}\right)^2 + \alpha|T|$$

      is as small as possible. Here, $|T|$ indicates the number of terminal nodes of the tree T and R_m is the region corresponding to the mth terminal node of the tree, and $\hat{y}_{R_m}$ is the predicted response associated with this region
    - As α increases, there is a price to pay for having a tree with many terminal nodes
- This entire algorithm is summarized below



Algorithm 8.1 *Building a Regression Tree*

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.
2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
3. Use K-fold cross-validation to choose α. That is, divide the training observations into K folds. For each $k = 1, \ldots, K$:

   (a) Repeat Steps 1 and 2 on all but the kth fold of the training data.

   (b) Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.

   Average the results for each value of α, and pick α to minimize the average error.
4. Return the subtree from Step 2 that corresponds to the chosen value of α.

Classification Trees

- The classification tree is used to predict a qualitative response instead of a quantitative one. Each observation is predicted to belong to the most commonly occurring class of observations in the region in which it belongs
- We use recursive binary splitting to grow classification trees where, instead of RSS, the assessed metric is classification error rate, which is the fraction of training observations in the region that do not belong to the most common class. This error rate is given by
  $$E = 1 - \max_k(\hat{p}_{mk})$$
  where $\hat{p}_{mk}$ is the proportion of training observations in the mth region that are from the kth class
- Alternative Measurements
  - The above error rate is not always the best for growing accurate trees
  - One alternative metric is the **Gini index** given by $G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$ which is a measure of total variance across all classes. The index will take a small value if all $\hat{p}_{mk}$ values are close to zero or one, so a small value indicates a regions purity, meaning it contains mostly observations of a single class
  - An alternative is the **entropy**, given by
    $$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$
    which operates like the Gini index
- These measurements can be used for tree creation and pruning, although classification error is preferable for pruning if prediction accuracy is the goal
- Trees can also consider qualitative predictors, having branches split to one or some of the classes

Using Decision Trees Over Other Models

- If the underlying structure of data is linear, a linear model typically performs better. However, tree models are more flexible and can compensate for complex variable relationships. Cross validation can be used to assess test errors and determine which methods are better
- Tree Pros
  - Are easy to explain and interpret
  - Mimic human decision making processes better than regression and classification
  - Can be explained and displayed graphically, even to non-experts
  - Can easily handle qualitative predictors without creation of dummy variables
- Tree Cons
  - Trees do not have the same level of prediction accuracy as other approaches
  - Trees are non-robust; small changes in data can cause large change in final estimated tree, meaning they have high variance
  - Ensemble methods that aggregate many decision trees can mitigate these issues

Bagging Ensemble Method

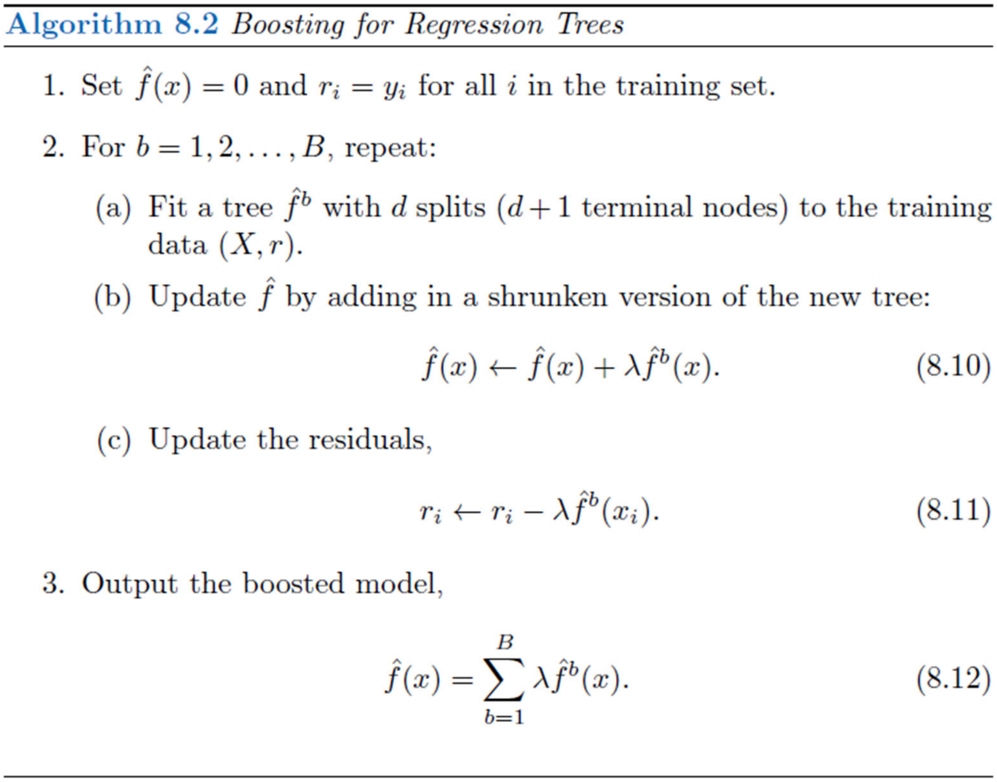
- To reduce the variance issues with trees, we employ bagging, a method that builds B models $\hat{f}^1(x), \hat{f}^2(x), \dots, \hat{f}^B(x)$ from B bootstrapped generated data sets to obtain a final averaged model given by $\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x)$
- To apply this to trees, we grow B regression trees that are not pruned. Hence, the trees have high variance but low bias. We then average the trees to reduce the variance. If B is sufficiently high, we can compensate for the errors of tree models
- To mimic this approach for classification trees, we record the class predicted by each of the B trees and take the majority vote to be predicted as the final class
- Out of Bag Error Estimation
  - When bagged trees are fit, the only use about $2/3$ of the given data due to the bootstrap resampling
  - We can use the $1/3$ of unused observations for that specific bagged tree to estimate the test error of the bagged tree
  - Additionally, since there are $B/3$ trees that were not trained using a given observation, we can run the given observation through these $B/3$ trees and average (or in the case of classification, majority vote) their test errors to get the estimated test error for that value. Aggregating all these test errors provides an estimate for the model's overall test error
- Variable Importance
  - It can be extremely difficult to interpret the results of bagging; bagging sacrifices interpretability for prediction accuracy
  - Yet, by recording how much the RSS or Gini index decreased by splits over each predictor, averaged over all B trees, we can indicate variable importance and track the most important variables

Random Forests Ensemble Method

- Random forests improves bagged trees by decorrelating the trees by instead of considering all predictors, only considering a random sample of them. Typically, this sample is of $m \approx \sqrt{p}$, meaning each individual tree is not allowed to consider the majority of the predictors
- This prevents massive correlation in the trees and the influence of a few salient predictors, situations that would otherwise negate any intended variance-reducing advantages. On average, a specified predictor is not considered by $\frac{p-m}{p}$ splits of the data
- Using a small m value will be helpful when there is a large amount of correlated predictors

Boosting Ensemble Method

- Boosting is similar to bagging except that each individual tree is grown sequentially, that is, using information from previously grown trees
- The Algorithm



Algorithm 8.2 *Boosting for Regression Trees*

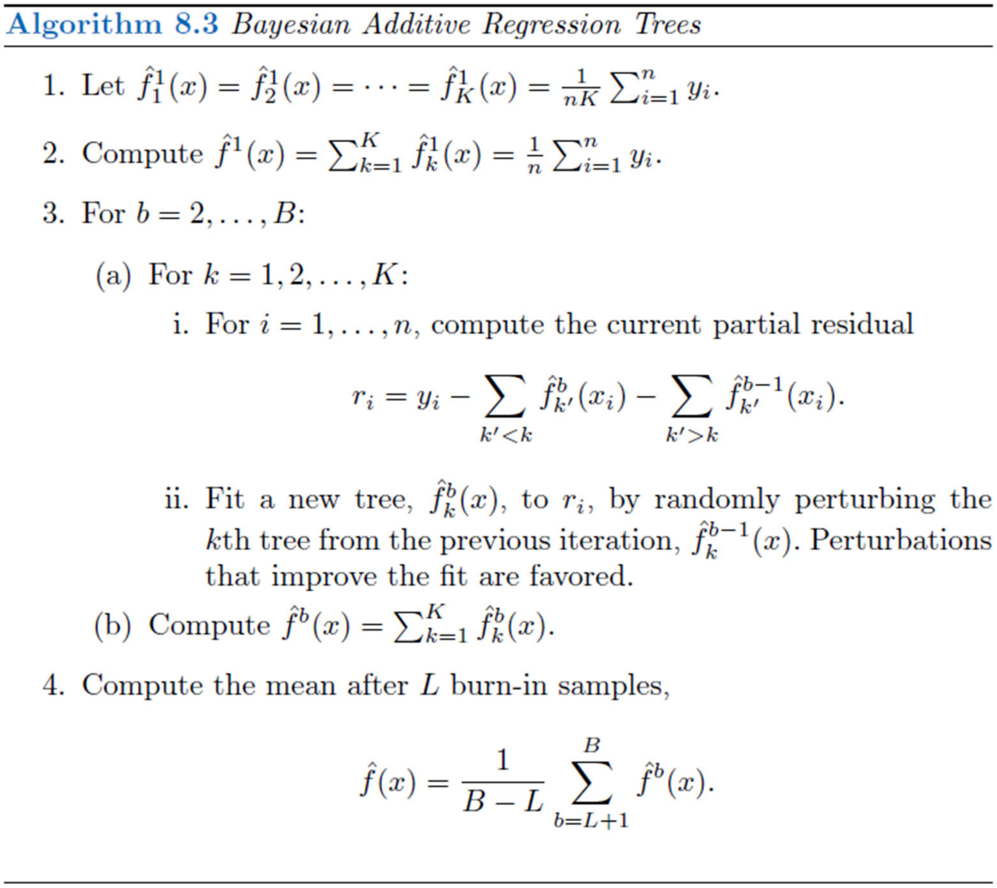
```
1. Set $\hat{f}(x) = 0$ and $r_i = y_i$ for all $i$ in the training set.
2. For $b = 1, 2, \ldots, B$, repeat:
    (a) Fit a tree $\hat{f}^b$ with $d$ splits ($d + 1$ terminal nodes) to the training data $(X, r)$.
    (b) Update $\hat{f}$ by adding in a shrunken version of the new tree:
            $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x).$    (8.10)
    (c) Update the residuals,
            $r_i \leftarrow r_i - \lambda \hat{f}^b(x_i).$    (8.11)
3. Output the boosted model,
            $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^b(x).$    (8.12)
```

  - The boosting approach learns slowly
  - Given the current model, a decision tree is fit to the residuals of this model instead of the response. This new decision tree is added into the fitted function in order to update the residuals, slowly improving the model in the areas where it does not preform well
  - Boosting has three tuning parameters
    - The number of trees B. If this parameter is too large, boosting can slowly overfit
    - The shrinkage parameter λ. This controls the rate at which boosting learns
    - The number of splits in the tree, d. This controls the complexity of the ensemble. A value $d = 1$ often works well, and is called a **stump**. More generally, d is the interaction depth since it controls how much the predictors can interact since the number of splits determines how many variables at most are involved. Smaller trees lead to more interpretable models

Bayesian Additive Regression Trees (BART) Ensemble Method

- BART uses approaches from bagging, random forests, and boosting. It builds trees in a random manner, like random forests and bagging, while also trying to capture something not captured by the current model, like boosting

- The Algorithm



Algorithm 8.3 *Bayesian Additive Regression Trees*

```
1. Let $\hat{f}_1^1(x) = \hat{f}_2^1(x) = \cdots = \hat{f}_K^1(x) = \frac{1}{nK}\sum_{i=1}^n y_i$.
2. Compute $\hat{f}^1(x) = \sum_{k=1}^K \hat{f}_k^1(x) = \frac{1}{n}\sum_{i=1}^n y_i$.
3. For $b = 2, \ldots, B$:
   (a) For $k = 1, 2, \ldots, K$:
       i. For $i = 1, \ldots, n$, compute the current partial residual
          $$r_i = y_i - \sum_{k'<k} \hat{f}_{k'}^b(x_i) - \sum_{k'>k} \hat{f}_{k'}^{b-1}(x_i).$$
       ii. Fit a new tree, $\hat{f}_k^b(x)$, to $r_i$, by randomly perturbing the
           $k$th tree from the previous iteration, $\hat{f}_k^{b-1}(x)$. Perturbations
           that improve the fit are favored.
   (b) Compute $\hat{f}^b(x) = \sum_{k=1}^K \hat{f}_k^b(x)$.
4. Compute the mean after $L$ burn-in samples,
   $$\hat{f}(x) = \frac{1}{B-L}\sum_{b=L+1}^{B} \hat{f}^b(x).$$
```

  - Let K denote the number of regression trees
  - Let B denote the number of iterations the BART algorithm is run
  - $\hat{f}_k^b(x)$ represents the prediction at x for the kth regression tree used in the bth iteration
  - At the end of each iteration, the K trees from that iteration are summed according to $\hat{f}^b(x) = \sum_{k=1}^K \hat{f}_k^b(x)$ for $b = 1, \ldots, B$
  - Since the first iterations are generally not that useful, we discard L of them in the burn-in period
  - A key element in step 3a is that we do not fit a fresh tree to the current partial residual; instead, we try to improve the fit to the current partial residual by modifying the tree obtained in the previous iteration
- BART can be viewed as a Bayesian approach to fitting an ensemble of trees since each time we randomly perturb a tree in order to fit the residuals
- We can also view the BART algorithm as a Markov chain Monte Carlo algorithm for fitting the model
- We typically choose large values for B and K with moderate values for L

Summarizing Ensemble Methods

- Bagging – Trees are grown independently on random samples of the observations, making them quite similar to each other, meaning bagging can get caught in local regions and fail to explore the whole model space
- Random forest – The trees are grown independently on random samples of the observations where each split on the tree is performed using a random subset of the features to decorrelating tress and leading to more thorough exploration of model space
- Boosting – We only use the original data and do not draw any random samples. The trees are grown using a slow learning approach where each new tree is fit to the signal that is left over from the earlier trees and then shrunken down before it is used
- BART – we only make use of the original data and the trees are grown successively with each new tree being perturbed in order to avoid local minima and achieve a better exploration of the model space