Stat 587 Outline

Steve Vardeman
Iowa State University

July 3, 2018

Abstract

This is an outline for (one version of) Stat 587 at Iowa State University. This basic course in the analysis of research data is intended for engineering, physical science, and mathematical sciences graduate students. It has as a **real** prerequisite an undergraduate course in applied statistics. (While the course is more or less "self-contained," its pace makes it impossible to grasp without this background.) It assumes a calculus III mathemtical background.

References in this outline are to *Basic Engineering Data Collection and Analysis* by Vardeman and Jobe (V&J), *Statistical Methods for Quality Assurance* by Vardeman and Jobe (V&J SMQA), and *An Introduction to Statistical Learning* by James, Witten, Hastie, and Tibshirani (JWH&T).

Contents

Part I

Review of Basic Probability and One-, Two-, and r-Sample Inference

1 Course Introduction and Probability Basics

(Text Reference/Reading: V&J Chapters 1 and 3, Appendix A.1)

Statistics is the study of how best to

- collect **data**,
- summarize **data**, and
- draw properly "hedged" conclusions (inferences) from **data**,

all in a context that recognizes the omnipresence of **variability** (and "randomness").

Probability is the mathematics intended to describe "chance" (or "randomness"). It is a worthy subject in its own right (providing important modeling for physical systems that are not "deterministic"/perfectly predictable). However, for Stat 587 purposes it is primarily a *tool* used in statistical analysis. Considerable overhead is involved in providing even the minimal probability background needed for statistical inference.

Dealing with data in any but the very simplest contexts also requires an appropriate **computational engine**. Stat 587 will use the open source `R` system and the `RStudio` interface both to do statistical computations from empirically derived datasets, *and* to do stochastic/probabilistic simulations (as a way of handling some probability calculations and of illustrating probability concepts).

Probability is like any other mathematical theory in that one begins with some notation and axioms and derives implications of these (theorems and the like). We'll use these implications to start with "some" probabilities and find others of interest that are consistent with the axioms and the input probability values.

We start with notation, set theory, and Boolean algebra concepts. Some basic notation for probability is in Table 1.

Table 1: Probability Notation

Typical Notation	**Other Notations**	**Meaning**
$\mathcal{S}$		a **sample space**/universe/universal set
$s \in \mathcal{S}$		an **outcome**/element of $\mathcal{S}$
$A \subset \mathcal{S}$		an **event**/set of outcomes of interest
A or B	$A \cup B$	the union of events A, B
A and B	$A \cap B$	the intersection of events A, B
not A	$A^c/\bar{A}$	the complement of the event A
$\emptyset$		the **empty event**/empty set

$\mathcal{S}$ can be thought of as a listing of "all things that might happen" in a "chance situation." Two events with no outcomes in common are called "mutually exclusive" events. In symbols, events A, B are mutuality exclusive when A and $B = \emptyset$.

The basic **axioms of probability** (rules of operation of a probability model) concern a function $P(A)$ that assigns numbers (probabilities) to events. These are

1. $0 \leq P(A) \leq 1$,
2. $P(\mathcal{S}) = 1$ (and in light of 3. below, $P(\emptyset) = 0$), and
3. for mutually exclusive events $A_1, A_2, \ldots$,

   $$P(A_1 \text{ or } A_2 \text{ or } \ldots) = P(A_1) + P(A_2) + \cdots$$

   (probabilities for mutually exclusive events add to make the probability that one of them occurs).

Probabilities are theoretical values meant to behave like empirical relative frequencies. Any system of numbers satisfying these axioms specifies a mathematically *valid* probability model. Whether that mathematically coherent model is useful or realistic in the physical world is a completely separate question that can be answered only through comparison of its predictions to empirical reality.

Some simple "theorems" (consequences of the basic axioms) of probability are the following.

Theorem 1 $P(\textit{not } A) = 1 - P(A)$

Theorem 2 ***(The "addition rule")*** $P(A \textit{ or } B) = P(A) + P(B) - P(A \textit{ and } B)$

Theorem 3 *If $\mathcal{S}$ is finite and outcomes are equally likely, then*

$$P(A) = \frac{\#(A)}{\#(\mathcal{S})}$$

2 Conditioning and Independence

(Text Reference/Reading: V&J Appendix A.1)

A basic notion of probability modeling is that of **conditional probability**. This represents what is appropriate as an assignment of chance given some partial information that makes a "reduced sample space" appropriate.

Definition 4 *If B is an event with $P(B) > 0$, the conditional probability of A given B is (the ratio of probabilities)*

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Simple multiplication through by $P(B)$ in Definition 4 produces a small theorem.

Theorem 5 ***(The "multiplication rule")*** $P(A \text{ and } B) = P(A|B)\,P(B)$

The possibility that in some cases $P(A|B)$ agrees exactly with $P(A)$ might be interpreted to mean that knowledge of the occurrence of B then has no impact on one's assessment of the likelihood of occurrence for A ... B is somehow uninformative concerning A. When $P(A|B) = P(A)$ the jargon "event A is **independent** of event B" is used. As it turns out, the four relationships

$$P(A|B) = P(A)\,,\ P(A|B^{\mathrm{c}}) = P(A)\,,\ P(B|A) = P(B)\,,\ \text{and } P(B|A^{\mathrm{c}}) = P(B)$$

are all equivalent (each implies all three others). This observation together with the fact that *under independence* (i.e. provided $P(A|B) = P(A)$)

$$P(A) = \frac{P(A \text{ and } B)}{P(B)}$$

and so

$$P(A)\,P(B) = P(A \text{ and } B)\,,$$

suggest a definition of **joint independence** of multiple events.

Definition 6 *Events $A_1, A_2, \ldots$ are* ***independent*** *provided that every event that is an intersection of two or more of these or their complements has probability that is the product of the corresponding probabilities.*

In the case of two events, independence (that is most easily understood as $P(A|B) = P(A)$) is equivalent to all of the relationships

$$\begin{gathered} P(A)\,P(B) = P(A \text{ and } B)\,, P(A)\,P(B^{\mathrm{c}}) = P(A \text{ and } B^{\mathrm{c}})\,, \\ P(A^{\mathrm{c}})\,P(B) = P(A^{\mathrm{c}} \text{ and } B)\,, \text{ and } P(A^{\mathrm{c}})\,P(B^{\mathrm{c}}) = P(A^{\mathrm{c}} \text{ and } B^{\mathrm{c}}) \end{gathered}$$

holding true. Independence means that the individual probabilities are all that is needed to specify all joint probabilities for the events.

3 Counting

(Text Reference/Reading: V&J Appendix A.3)

This material is not statistics nor even really probability, but rather some simple discrete mathematics. But it is useful in application of Theorem 3. Both for that reason and because it is helpful in understanding the form of a so-called "binomial distribution" useful in inference for proportions, we include it here. There is one basic idea/principle and two important implications to understand.

A counting principle: If a complex action can be accomplished in a series of r steps, the first of which can be accomplished in n_1 ways, the second that can subsequently be accomplished in n_2 ways, ... , and the rth that can subsequently be accomplished in n_r ways, then the entire action can be accomplished in

$$n = n_1 \cdot n_2 \cdot \cdots \cdot n_r$$

ways.

This basic principle leads directly to the solution of two generic counting problems and corresponding sets of jargon and notation.

A first generic counting problem: The number of ordered lists possible for r out of n distinguishable items (no repetitions allowed) is

$$P_{n,r} = n\,(n-1)\,(n-2)\cdots(n-r+1) = \frac{n!}{(n-r)!}$$

(usually called "the number of **permutations** of n things taken r at a time").

Notice that in making an ordered list of r out of n items, one might think of *first* choosing r items from n and *then* ordering them. That clearly implies that the number of ways that the items can first be chosen is the ratio of $P_{n,r}$ to $P_{r,r}$. This leads to the second generic problem/result.

A second generic counting problem: The number of unordered collections of r out of n distinguishable items is

$$\binom{n}{r} = \frac{P_{n,r}}{P_{r,r}} = \frac{n!}{(n-r)!r!}$$

(usually called "the number of **combinations** of n things taken r at a time" or sometimes "n choose r").

4 Random Variables and Generic Discrete Distributions

(Text Reference/Reading: V&J Section 5.1)

"Chance" situations lead to quantities whose values might be described as subject to "random" influences. These are known as **random variables** in probability modeling. Prior to observation, these can be described in terms of probabilities. Typical elementary symbology is to use capital Roman letters near the end of the alphabet (like X, Y, or Z) to stand for such objects.

Modeling for random variables is done in exact analogy to description of coordinate variables in the mechanics of mass distributions. That is, **probability distributions** describe how probability is spread out in 1-d, or 2-d, or ... (depending upon how many random variables are under discussion) using tools exactly parallel to those used to describe how **mass distributions** spread mass around on a line, in a plane, etc. Just as in mechanics, the tools for describing discrete distributions are not exactly the same as those used in describing their (idealized) continuous counterparts. The former involve discrete mathematics (technically only requiring the use of algebra) while the latter involve continuous mathematics (calculus of 1 or 2 or ... variables, depending upon the dimensionality of the problem under study). We begin with discussion of modeling for *single*/individual *discrete* random variables.

Discrete probability models for single (1-d) random variables use a sample space $\mathcal{S}$ specifying outcomes for the random variable that is either finite or (at least) "countable" in the sense that it is like the integers or non-negative integers. In this context, probabilities for a random variable X can be specified by a so-called **probability mass function**, $f(x)$, giving for each value of x the probability that X takes that value. That is, a discrete **pmf** (probability mass function) is

$$f(x) = P[X = x]$$

read "$f(x)$ is the probability that the random variable X takes the value x."

To describe/provide summaries of discrete distributions (equally, their pmfs) one can make spike graphs or probability histograms and compute analogues of the "moments" of mass distributions in mechanics.

Definition 7 *The* ***expected or mean value of the discrete random variable*** X *(the mean of its distribution) is*

$$EX = \sum_x x f(x)$$

and alternative notation for this is μ_X*.*

The mean of a discrete probability distribution is exactly the *center of mass* of that 1-d distribution from mechanics. (The fact that $\sum_x f(x) = 1$ says that the usual divisor appearing in a center of mass formula is not needed in the present situation. A probability distribution is a mass distribution with total mass 1.) The concept of a mass moment of inertia (around the center of mass) in mechanics has an analogue in probability as the variance of a random variable.

Definition 8 *The* ***variance of the discrete random variable*** X *(the variance of its distribution) is*

$$VarX = \sum_x (x - EX)^2 f(x) = \sum_x (x - \mu_X)^2 f(x)$$

and alternative notation for this is σ_X^2*.*

The variance of a discrete probability distribution is a mean squared deviation from the average/expected value of X, a measure of spread for the distribution. It has units that are the squares of the original units (the units of X). A related measure of distributional spread that has the *same* units as X is the standard deviation.

Definition 9 *The* ***standard deviation of a random variable*** X *(the standard deviation of its distribution) is*

$$\sigma_X = \sqrt{VarX}$$

The standard deviation is a root mean squared deviation from the average/expected value of X.

The variance of X is an average of the random variable $Y = (X - \mu_X)^2$ (that is a function of X). The are many other cases where it is useful to employ the concept of the mean of an arbitrary function of X, say $h(X)$. One could (at least in principle) work out the distribution for $Y = h(X)$ (find the possible values and corresponding probabilities for Y) in order to then find the mean of Y. Another way of proceeding (that turns out to be equivalent) is to simply define a mean or expected value for $h(X)$ in terms of the distribution for X.

Definition 10 *The expected value or mean of* $h(X)$ *for a discrete random variable* X *is*

$$Eh(X) = \sum_x h(x) f(x)$$

An alternative (to the pmf $f(x)$) way to specify the distribution of a random variable X (discrete or not) is through a so-called cumulative distribution function. This is a function giving probabilities for X taking a value in intervals $(-\infty, x]$.

Definition 11 *The* ***cumulative distribution function for the random variable*** X *(the cdf of its distribution) is*

$$F(x) = P[X \leq x]$$

In the case of discrete variables, cdfs are stair-step functions, increasing left to right, jumping up the amount $f(x)$ at the value x.

5 Standard Discrete Distributions

(Text Reference/Reading: V&J Section 5.1)

There are a number of standard discrete probability distributions. In Stat 587 we'll consider 3 of them. Two are distributions related to sequences of "success/failure trials." A third is a model for the number of occurrences of a relatively rare phenomenon across a fixed "interval" of time or space. (This latter actually also has a connection to success/failure trials, but the connection is not so direct as for the first two distributions.)

So, to begin, we consider "trials" $1, 2, 3, \ldots$ that each will yield one of two possible outcomes. We'll arbitrarily call one of those two possible outcomes a "success" (S) and the other a "failure" (F) (without attaching any positive or negative connotations to these labels). In this context the two counting variables

$$X = \text{the number of S's in the first } n \text{ trials} \tag{1}$$

and

$$Y = \text{the index of the trial on which the first S occurs} \tag{2}$$

are often of interest. Under so-called "Bernoulli process" assumptions, it is possible to identify simple formulas for pmfs for them.

A **Bernoulli process model** for S/F trials is one where

1. trials are independent in the sense that the events

   $$A_i = \text{trial } i \text{ yields a success}$$

   are independent events, and

2. the probability of success on each trial is p (a constant value) across all trials.

Under a Bernoulli process model, the distribution for the variable X in display (1) has pmf

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

This is the so-called **binomial** pmf with parameters (n, p) . (The name derives from the fact that its values are the terms in a binomial expansion of $1 = (p + (1-p))^n$.) *For this distribution* there are very simple forms for the mean and variance. That is, for $X \sim \text{Bin}(n, p)$ (read "X distributed as binomial (n, p)"),

$$\text{E}X = np \quad \text{and} \quad \text{Var}X = np(1-p)$$

(that is $\sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} = np$ and $\sum_{x=0}^{n} (x - np)^2 \binom{n}{x} p^x (1-p)^{n-x} = np(1-p)$).

Under a Bernoulli process model, the variable Y in display (2) has pmf

$$f(y) = \begin{cases} p(1-p)^{y-1} & y = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases}$$

This is the so-called **geometric** pmf with parameter p. (The name derives from the fact that its values are values in a geometric infinite series that adds to 1.) *For this distribution* there are very simple forms for the mean and variance. That is, for $Y\sim$Geo(p),

$$\mathrm{E}Y=\frac{1}{p}\ \text{ and }\ \mathrm{Var}Y=\frac{(1-p)}{p^2}$$

Further, cumulative probabilities are easily computed from the fact that for $y=1,2,\ldots$

$$1-F(y)=P[Y>y]=p^y$$

so that for such y,

$$F(y)=1-p^y$$

The so-called **Poisson** probability distributions are used to model variables W that are counts of the number of occurrences of a relatively "rare" phenomenon across a specified interval of time or space. (Standard examples are numbers of detectable cracks in a fixed surface area of a material specimen, numbers of information packets arriving at a switching center during a fixed time interval, etc.) Under assumptions that

1. numbers of occurrences in non-overlapping intervals are independent,
2. the probability of a single occurrence in a small interval is approximately proportional to the size of the interval, and
3. relative to the probability of a single occurrence in a small interval, the probability of more than one occurrence is negligible,

it is possible to derive the Poisson pmf with parameter λ

$$f(w)=\begin{cases}\dfrac{\lambda^w\exp(-\lambda)}{w!} & \text{for } w=0,1,2,\ldots\\[2ex] 0 & \text{otherwise}\end{cases}$$

for the overall count, W. As it turns out, the mean and variance of this distribution are both λ, i.e.

$$\mathrm{E}W=\lambda\ \text{ and }\ \mathrm{Var}W=\lambda$$

(that is, $\sum_{w=0}^{\infty}w\dfrac{\lambda^w\exp(-\lambda)}{w!}=\lambda$ and $\sum_{w=0}^{\infty}(w-\lambda)^2\dfrac{\lambda^w\exp(-\lambda)}{w!}=\lambda$).

6 Generic Continuous Distributions

(Text Reference/Reading: V&J Section 5.2)

At least as a mathematically convenient idealization, it is common to consider continuous models for random variables. These are the probability analogues of the continuous mass distributions of mechanics. Just as a continuous mass distribution is specified in terms of a mass density that is integrated over appropriate values to find the mass in the region of integration, a continuous probability distribution is specified in terms of a probability density that is integrated over appropriate values to find probabilities of interest. That is, in 1-d (for a single random variable) we have the following.

Definition 12 *A random variable X is said to have a* ***continuous*** *distribution provided there is a function $f(x)$ (called a* ***probability density function****) such that for any $a < b$*

$$P\left[a < X < b\right] = \int_a^b f(x)\,dx$$

It follows from this definition that a **pdf** (probability density function) f for X is related to the cdf for X by

$$F(x) = P\left[X \le x\right] = \int_{-\infty}^{x} f(t)\,dt$$

(and then $f(x) = \frac{d}{dx}F(x)$).

Moments for discrete mass distributions have analogues for continuous probability distributions.

Definition 13 *The* ***expected or mean value of the continuous random variable*** *X (the mean of its distribution) is*

$$EX = \int_{-\infty}^{\infty} x f(x)\,dx$$

and alternative notation for this is μ_X.

Definition 14 *The* ***variance of the continuous random variable*** *X (the variance of its distribution) is*

$$VarX = \int_{-\infty}^{\infty} (x - EX)^2 f(x)\,dx = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\,dx$$

and alternative notation for this is σ_X^2.

The mean and variance of a continuous probability distribution have the same interpretations (as center of mass and measure of spread of the distribution) as were offered for discrete ones. Definition 9 (that defines the standard deviation as the square root of the variance) applies equally to discrete and continuous variables.

The expected value of $h(X)$ for a continuous X is analogous to that for the discrete case.

Definition 15 *The expected value or mean of $h(X)$ for a continuous random variable X is*

$$Eh(X) = \int_{-\infty}^{\infty} h(x) f(x)\,dx$$

7 Standard Continuous Distributions

(Text Reference/Reading: V&J Section 5.2)

Just as there are useful standard discrete distributions, there are standard pdf forms that prove useful in many applications. We here we will consider a few of these.

Possibly the simplest continuous distributions are those that are uniform on some interval. That is, for $\theta_1 < \theta_2$ the pdf

$$f(x) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & \text{otherwise} \end{cases}$$

specifies the so-called **Uniform** (θ_1, θ_2) distribution. The most common version of this is the case where $\theta_1 = 0$ and $\theta_2 = 1$. This case is the target distribution for standard "random number generators" and is the fundamental building block of modern stochastic/probabilistic simulation.

The cdf for a Uniform (θ_1, θ_2) random variable is

$$F(x) = \begin{cases} 0 & x < \theta_1 \\ \dfrac{x - \theta_1}{\theta_2 - \theta_1} & \theta_1 \le x \le \theta_2 \\ 1 & x > \theta_2 \end{cases}$$

The mean and variance for a uniform distribution are relatively simple and intuitively reasonable. That is, if $X \sim \mathrm{U}(\theta_1, \theta_2)$

$$\mathrm{E}X = \frac{\theta_1 + \theta_2}{2} \quad \text{and} \quad \mathrm{Var}X = \frac{1}{12}\left(\theta_2 - \theta_1\right)^2$$

The "**normal**" or **Gaussian** distributions are the archetypal "bell-shaped" distributions. The Gaussian pdf with parameters μ and σ^2 is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x - \mu\right)^2\right)$$

As it then turns out

$$\mathrm{E}X \equiv \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x - \mu\right)^2\right) dx = \mu$$

and

$$\mathrm{Var}X \equiv \int_{-\infty}^{\infty} \left(x - \mu\right)^2 \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(x - \mu\right)^2\right) dx = \sigma^2$$

That is, the parameters μ and σ^2 are in fact the mean and variance of the distribution.

The case of the Gaussian distribution with $\mu = 0$ and $\sigma = 1$ is called the **standard normal** case. For this case, the special notation

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

is used for the pdf, and the standard normal cdf

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)\, dt$$

is routinely tabled. Probabilities for *any* normal distribution can be obtained using it. That is, for $X \sim \mathrm{N}(\mu, \sigma^2)$ and $a < b$,

$$\begin{aligned} P[a < X < b] &= \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx \\ &= \int_{\frac{a-\mu}{\sigma}}^{\frac{b-\mu}{\sigma}} \phi(z)\, dz \\ &= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \\ &= P\left[\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right] \end{aligned}$$

That is, for $X \sim \mathrm{N}(\mu, \sigma^2)$,

$$Z = \frac{X-\mu}{\sigma}$$

is standard normal.

Another kind of continuous distribution frequently used in engineering and physical science applications is the family of **Weibull** distributions and the sub-family of so-called **exponential** distributions. For parameters $\alpha > 0$ and $\beta > 0$, the distribution (putting all its probability on $(0, \infty)$) with cdf

$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right) & \text{for } x \geq 0 \end{cases}$$

is the Weibull(α, β) distribution. The corresponding pdf can be found by differentiating this fairly simply cdf. The result is that $X \sim$Weibull(α, β) has pdf

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ \frac{\beta}{\alpha^{\beta}} x^{\beta-1} \exp\left(-\left(\frac{x}{\alpha}\right)^{\beta}\right) & \text{for } x > 0 \end{cases}$$

The parameter β controls the shape of this pdf, while the parameter α controls the scale. While the mean and variance for the distribution are not impossible to work out, they are not particularly simple. For example

$$\mathrm{E}X = \alpha\Gamma\left(1 + \frac{1}{\beta}\right)$$

in terms of the "special function" Γ. Something that *is* fairly simple to find is the median of the Weibull(α, β) distribution. Setting $F(x) = .5$ and solving for x reveals that the "50% point" of the distribution is

$$median = \alpha \exp\left(-\frac{.3665}{\beta}\right)$$

An important special case of the Weibull family is that where $\beta = 1$. This is the case of the **exponential** distributions, where

$$F\left(x\right) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - \exp\left(-\left(\dfrac{x}{\alpha}\right)\right) & \text{for } x \geq 0 \end{cases}$$

and

$$f\left(x\right) = \begin{cases} 0 & \text{for } x < 0 \\ \dfrac{1}{\alpha}\exp\left(-\left(\dfrac{x}{\alpha}\right)\right) & \text{for } x > 0 \end{cases}$$

Then if X ~Exp(α)

$$\mathrm{E}X = \alpha \text{ and } \mathrm{Var}X = \alpha^2$$

8 Joint Distributions of Several Random Variables

(Text Reference/Reading: V&J Section 5.4)

Most applications of probability (particularly in statistics) involve more than a single random variable. The tools discussed thus far must be extended in order to describe the joint behavior of these multiple variables. For example, for two random variables X and Y, one needs ways of specifying quantities like

$$P[X > Y]$$

So we consider some simple parts of the theory and use of multivariate distributions. Continuing the analogy with mechanics, this material is parallel to specification and use of mass distributions of mechanics in more than a single dimension. For simplicity of exposition, this discussion will be carried out primarily for the case of bivariate distributions (joint distributions of pairs of random variables) parallel to 2-dimensional mass distributions. The reader will need to reason by analogy to the generalization of this material to joint distributions of multiple (more than 2) random variables.

A jointly discrete distribution for two random variables X and Y is specified by a **joint probability mass function**, $f(x, y)$, giving for each pair of values (x, y) the probability that the random *vector* (X, Y) takes that *pair of values.* That is, a discrete **joint pmf** (joint probability mass function) is

$$f(x, y) = P[X = x \text{ and } Y = y]$$

and the notation is read "$f(x, y)$ is the probability that the random vector (X, Y) takes the value (x, y)." A joint pmf can be represented by either a formula giving its values, or in the case that the set of (x, y) pairs receiving positive probability is finite, a two-way table giving those probabilities. For $\mathcal{R}$ a subset of 2-dimensional space $\Re^2$, the probability associated with jointly discrete variables (X, Y) taking a (vector) value (x, y) in $\mathcal{R}$ is computed by simply summing values of $f(x, y)$, That is

$$P[(X, Y) \in \mathcal{R}] = \sum_{(x,y) \in \mathcal{R}} f(x, y)$$

Associated with a joint distribution for a pair of random variables (X, Y) are the individual distributions of the variables (considered one at a time). These individual distributions are called **marginal distributions**. For jointly discrete (X, Y), the marginal distributions are discrete and have pmfs that can be easily derived from the joint pmf by simple addition. These are

$$f_X(x) = \sum_y f(x, y) \quad \text{and} \quad f_Y(y) = \sum_x f(x, y)$$

The appropriateness of the language "marginal" distribution is especially evident in the discrete case from the fact that these marginal pmfs can be thought of as given by row or column sums across tables of joint probabilities (and recorded in the "margins" of the table).

Expected (mean) values for functions of jointly discrete random pairs (X, Y) are defined in exact analogy to the univariate case.

Definition 16 *The expected value or mean of $h(X, Y)$ for a jointly discrete random pair (X, Y) is*

$$Eh(X, Y) = \sum_{(x,y)} h(x, y) f(x, y)$$

A jointly continuous distribution for two random variables X and Y is specified by a **joint probability density function**, $f(x,y)$. (This is the analogue of a function specifying mass density in the x-y plane.) Probabilities are found by (bivariate) integration over a region of interest. That is, for $\mathcal{R}$ a subset of 2-dimensional space $\Re^2$, the probability associated with jointly continuous variables (X,Y) taking a (vector) value (x,y) in $\mathcal{R}$ is computed as

$$P[(X,Y)\in\mathcal{R}]=\int\int_{\mathcal{R}} f(x,y)\,dxdy$$

For jointly continuous (X,Y), the marginal distributions are continuous and have pdfs that are derived from the joint pdf by integration. These are

$$f_X(x)=\int f(x,y)\,dy \quad\text{and}\quad f_Y(y)=\int f(x,y)\,dx$$

Unsurprisingly, mean values for functions of jointly continuous random pairs (X,Y) are defined as integrals.

Definition 17 *The expected value or mean of $h(X,Y)$ for a jointly continuous random pair (X,Y) is*

$$Eh(X,Y)=\int\int h(x,y)\,f(x,y)\,dxdy$$

9 Conditional Distributions and Independence of Random Variables

(Text Reference/Reading: V&J Section 5.4 and 5.5)

Associated with a joint distribution for a pair of random variables (X, Y) are distributions of the variables conditioned on the value of the other variable (distributions of one variable holding the value of the second fixed at a particular value of interest). These are (not surprisingly) called **conditional distributions**. For jointly discrete (X, Y), the conditional distributions are discrete and have pmfs that can be easily derived from the joint pmf by simply "renormalzing" a "slice" of the joint pmf by dividing by the value of the corresponding marginal pmf (i.e. using a row or column of a table specifying a joint pmf divided by the corresponding row or column sum of entries). That is,

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} \quad \text{and} \quad f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} \tag{3}$$

Jointly continuous distributions have continuous conditional distributions. For these, a conditional pdf is simply a "slice" of a joint pmf "renormalized" by dividing by the corresponding marginal pdf (the integral of that slice). Reinterpreting symbols for pmfs as symbols for pdfs, formula (3) serves not only to specify conditional pmfs, but conditional pdfs as well.

A particularly simple and easy-to-work-with situation is one where conditional distributions for a variable are all the same (regardless of the value of conditioning variable or variables). This can be thought of as modeling a circumstance where knowledge of the value of X provides no modification of one's thinking about Y (and vice versa) and is called **independence** of the random variables. In this case conditional distributions are, in fact, the marginal distributions (so functions $f_{X|Y}(x|y) = f_X(x)$ for all y and $f_{Y|X}(y|x) = f_Y(y)$ for all x). This means that for all (x, y)

$$f(x,y) = f_X(x) f_Y(y) \tag{4}$$

(the joint pmf or pdf is the product of the two marginal pmfs or pdfs).

Where k random variables $X, Y, \ldots, Z$ are being modeled as either jointly discrete or jointly continuous, the generalization of relationship (4) is that joint and marginal pmfs or pdfs are related by

$$f(x, y, \ldots, z) = f_X(x) f_Y(y) \cdots f_Z(z) \tag{5}$$

Where this relationship between joint and marginal distributions holds, the variables $X, Y, \ldots, Z$ are **independent**. (NOTICE that relationship (5) must hold for all k-tuples of inputs $(x, y, \ldots, z)$.)

Modeling (5) has many useful implications. For one, since modern simulation software aims to generate (pseudo-) random values that "look" independent, it is often easy to simulate a large number of realizations of $(X, Y, \ldots, Z)$ and plug them into a function $g(x, y, \ldots, z)$ and thereby simulate realizations of

$$U \equiv g(X, Y, \ldots, Z)$$

The empirical properties of these realizations can in turn be used to get approximate answers to probability problems involving U. One application of this idea particularly useful in engineering and physical science is that of making "propagation of error" analyses to understand how variation or uncertainty in inputs $X, Y, \ldots, Z$ propagates to the output U. (This is, e.g., often quite helpful in the analysis of measurement systems.)

Joint distributions of independence lead to simple formulas for means and variances for variables made up as linear combinations from them. That is, when k random variables $X, Y, \ldots, Z$ are independent and $a_0, a_1, a_2, \ldots, a_k$ are constants, the linear combination

$$U = a_0 + a_1 X + a_2 Y + \cdots + a_k Z$$

has mean

$$\mathrm{E}U = a_0 + a_1 \mathrm{E}X + a_2 \mathrm{E}Y + \cdots + a_k \mathrm{E}Z$$

(in other notation, $\mu_U = a_0 + a_1 \mu_X + a_2 \mu_Y + \cdots + a_k \mu_Z$) and variance

$$\mathrm{Var}U = a_1^2 \mathrm{Var}X + a_2^2 \mathrm{Var}Y + \cdots + a_k^2 \mathrm{Var}Z$$

(in other notation, $\sigma_U^2 = a_1^2 \sigma_X^2 + a_2^2 \sigma_Y^2 + \cdots + a_k^2 \sigma_Z^2$).

The so-called "propagation of error formulas" (see Section 5.5.4 of V&J) provide approximate means and variances for variables $U = g(X, Y, \ldots, Z)$ with general (non-linear) g and independent inputs $X, Y, \ldots, Z$ based on first order Taylor approximation of a function. It's easy enough to simulate values U, that in a world where computing power is plentiful, one might as well do so (rather than make the potentially far less accurate calculus-based approximations).

10 IID Models and the "Central Limit Effect"

(Text Reference/Reading: V&J Section 5.5)

A very important version of the use of independence in modeling multiple random variables, is that where each variable has the same marginal distribution. Such models are often termed "**iid**" (independent identically distributed) **models**. These are models for "random draws from a fixed (conceptually) infinite universe (or population)." They are used in engineering and physical science to describe

1. observation of a physically stable process, and
2. observation of purposely "random" sampling from a huge (relative to sample size) group of objects.

In such contexts, one might well want to suppose that n random variables $X_1, X_2, \ldots, X_n$ are independent with a common marginal distribution. (In statistical inference, this is the so-called "one sample model.")

One particularly important variable that can be made from such $X_1, X_2, \ldots, X_n$ is their sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$$

Since this is a linear combination of independent random variables, its theoretical mean and variance follow immediately from the formulas just presented. That is,

$$\mathrm{E}\bar{X} = \mu_{\bar{X}} = \frac{1}{n}\mathrm{E}X_1 + \frac{1}{n}\mathrm{E}X_2 + \cdots + \frac{1}{n}\mathrm{E}X_n = \mu = \mathrm{E}X$$

where $\mu = EX$ is standing for the mean of the common marginal probability distribution of $X_1, X_2, \ldots, X_n$. The random variable that is the arithmetic average of the sample has a mean/expected value that is the same as that of the marginal distribution. In a similar way,

$$\mathrm{Var}\bar{X} = \sigma^2_{\bar{X}} = \left(\frac{1}{n}\right)^2 \mathrm{Var}X_1 + \left(\frac{1}{n}\right)^2 \mathrm{Var}X_2 + \cdots + \left(\frac{1}{n}\right)^2 \mathrm{Var}X_n = \frac{1}{n}\mathrm{Var}X = \frac{1}{n}\sigma^2$$

where $\sigma^2 = VarX$ is standing for the variance of the common marginal probability distribution of $X_1, X_2, \ldots, X_n$. The random variable that is the arithmetic average of the sample has a variance that is that of the marginal distribution divided by the sample size. That is, in this context

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

There are two other probability facts that tell one even more about the distribution of $\bar{X}$ in an iid model. We'll state them both as theorems.

Theorem 18 *In an iid model, if the common marginal distribution of the variables $X_1, X_2, \ldots, X_n$ is normal, so also is the distribution of $\bar{X}$.*

Theorem 19 ***(The Central Limit Theorem)*** *In an iid model, if the common marginal distribution of the $X_1, X_2, \ldots, X_n$ has a finite variance and n is large, then the distribution of $\bar{X}$ is approximately normal in the sense that*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

is approximately standard normal.

The Central Limit Theorem promises that "for large n" probabilities for Z are approximately standard normal probabilities. The quality of those approximations of course increases with n.

11 Large Sample Confidence Limits for a Single Mean μ

(Text Reference/Reading: V&J Section 6.1)

One of the most basic questions of statistical inference is this:

> Based on observations $x_1, x_2, \ldots, x_n$ from a single population/universe/process, how might one make an estimate of the mean of that population/universe/process and attach to it a sensible quantification of reliability of that estimate?

The material of the last section begins to provide an answer to this question. We use it here to begin study of probability-based statistical inference. As a matter of notation for all that follows, (as is completely standard) we now *drop the convention* that random variables are always represented by capital letters, and alert the reader that it will be necessary to determine from context whether a letter is standing for a random variable, one of its possible values, or some constant.

For iid observations (from a stable process or large fixed population) with mean μ and standard deviation σ,

$$x_1, x_2, \ldots, x_n \quad ,$$

the previous section said that provided the distribution sampled is normal or the sample size (n) is large, the sample mean

$$\bar{x}$$

is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. This implies for example that (since for Z standard normal $P\left[-1.645 < Z < 1.645\right] = .90$),

$$P\left[\bar{x} \text{ is within } 1.645\frac{\sigma}{\sqrt{n}} \text{ of } \mu\right] \approx .90$$

But the event

$$\bar{x} \text{ is within } 1.645\frac{\sigma}{\sqrt{n}} \text{ of } \mu$$

is the event

$$\bar{x} - 1.645\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.645\frac{\sigma}{\sqrt{n}}$$

and so (before observations are made) the random limits

$$\bar{x} \pm 1.645\frac{\sigma}{\sqrt{n}}$$

have a 90% chance of bracketing μ. These might thus be called "90% (two-sided) confidence limits" for μ.

More generally, (*typically unusable* since σ will rarely be known when μ is to be estimated) **confidence limits** for μ are

$$\bar{x} \pm z\frac{\sigma}{\sqrt{n}} \tag{6}$$

Different choices of $z > 0$ produce different **confidence levels** $P\left[|Z| < z\right]$ for a two-sided interval with endpoints (6). One or the other of endpoints (6) can be used to make a one-sided interval for μ with confidence level $P\left[Z < z\right]$. The biggest drawback of this development is that formulas (6) involve the typically unknown σ. This has remedies, and the simplest can be used when n is large and is presented next.

It turns out that there is an extension of the central limit theorem says that under an iid model, for large n the variable

$$Z = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

(that involves the sample standard deviation s in place of the population standard deviation σ appearing in Theorem 19) is also approximately standard normal. The same logic that leads to limits (6) then leads to practical **confidence limits** for μ

$$\bar{x} \pm z \frac{s}{\sqrt{n}} \tag{7}$$

that do not involve the typically unknown population standard deviation. As is the case for limits (6), both of limits (7) can be used at once to make a two-sided interval and a single one of them can be used to make a one-sided interval (to make a lower or upper confidence bound).

It is essential to understand the sense in which there is a stated confidence associated with an interval made using endpoints like (7). A confidence level is a kind of "reliability" *of the inference method*, a "lifetime winning percentage" one would experience using the method repeatedly (sometimes having a good result and sometimes not). The reader should carefully study pages 341-342 of V&J in this regard and that full discussion will not be repeated here.

The "plus or minus part" of limits (6) or (7), namely

$$z \frac{\sigma}{\sqrt{n}} \text{ or } z \frac{s}{\sqrt{n}} \quad ,$$

might be termed a "margin of error" associated with estimating μ. Armed with a target (say m) for such a margin of error and values for the population standard deviation and confidence level (and therefore z), the equation

$$m = z \frac{\sigma}{\sqrt{n}}$$

can be solved for a sample size, n, producing that margin of error. This provides some elementary guidance for the "sample size question" (for estimating μ).

12 Large Sample Significance Testing for a Single Mean μ

(Text Reference/Reading: V&J Section 6.2)

A second standard type of probability-based statistical inference is called **hypothesis testing**. It has "significance testing" and "decision-making" forms. For reasons laid out in detail in V&J Section 6.2, Vardeman holds that the making of confidence limits is far more informative and practically important than hypothesis testing. However, because testing is common in the engineering and scientific literature that graduates of Stat 587 must read, it is necessary to also discuss it. The primary thrust of the V&J discussion of testing concerns the "significance testing" version of the methodology, but some attention is given to the decision-making version. It is introduced in Stat 587 in the context of large sample inference for a single mean.

Significance testing is essentially a methodology for *probabilistically assessing the strength of evidence in a dataset against the possibility that a given statement about model/population parameters is true.* It is a way of assessing whether one has enough data to clearly "see" the difference between a model parameter and some hypothesized numerical value for that parameter.

The 5-step significance testing format used in V&J provides a consistent way of presenting the results of significance tests, and its use will be required in Stat 587. The steps are these:

1. State a **null hypothesis**. In the simplest cases, this is a statement of the form

$$\mathrm{H}_0\text{:}\textit{parameter} = \#$$

   that (for a number of interest, #) embodies the "status quo"/"no data" view of the scenario under study.

2. State an **alternative hypothesis** of one of the three forms

$$\mathrm{H_a}\text{:}\textit{parameter} \begin{array}{c} > \\ \neq \\ < \end{array} \#$$

   that is meant to describe departures from H_0 that are of interest/that one wishes to be able to detect.

3. Give

   (a) (only) a (formula for a) **test statistic** (a data summary) to be used (NOT plugging data into the formula at this stage),

   (b) a complete specification (name and appropriate parameter values) of a probability distribution (a "null" or "reference" distribution) that describes variation in the test statistic if in fact the null hypothesis is exactly true, and

   (c) a specification of what type(s) values of the test statistic will be counted as evidence against H_0 and in favor of $\mathrm{H_a}$.

4. Compute the observed value of the test statistic. (This is where data are plugged into the formula from 3(a) and a single value corresponding to the sample is computed and displayed.)

5. Find, report, and interpret a *p*-**value** (or so-called "**observed level of significance**"). This is the probability that the reference distribution (in 3(b)) assigns to values of the test statistic more extreme (per 3(c)) than the one observed. *Small* *p*-values are counted as *evidence against* H_0 and *in favor of* H_a. They more or less indicate that one has enough data to see difference between "*parameter*" and "#". (However, a *p*-value MAY NOT be interpreted as a "probability that H_0 is true," a quantity that is simply without rational definition.)

The first application of the significance testing logic met in Stat 587 concerns large n tests of $\mathrm{H}_0{:}\mu = \#$ (based on an iid model for sampling a stable process or fixed large population) where a number appropriate in the applied context replaces #. The corresponding possible alternative hypotheses are

$$\mathrm{H}_\mathrm{a}{:}\mu > \#,\ \ \mathrm{H}_\mathrm{a}{:}\mu \neq \#,\ \text{ and }\ \mathrm{H}_\mathrm{a}{:}\mu < \#$$

The test statistic

$$Z = \frac{\bar{x} - \#}{\frac{\sigma}{\sqrt{n}}}$$

and its more typically relevant version (not involving the typically unknown σ)

$$Z = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}}$$

have approximately standard normal distributions when H_0 is true (and $\mu = \#$). Corresponding to the three possible alternative hypotheses are specifications 3(c) of observed values z (of the random variable Z) with respectively

$$\text{large } z,\ \text{ large } |z|\,,\ \text{ and small (large negative) } z$$

producing *p*-values respectively

$$1 - \Phi(z)\,,\ \ 2(1 - \Phi(|z|))\,,\ \text{ and }\ \Phi(z)$$

The decision-making version of hypothesis testing uses the observed value of a test statistic to choose between remaining with a "status quo" null hypothesis and being compelled to reject it in favor of the alternative hypothesis in light of the evidence provided by the data. There is standard jargon associated with this approach. Part of it is summarized in Table 2.

Table 2: Standard Testing Jargon

		Decision in Favor of	
		H_0	H_a
Actual	H_0		Type I Error
Situation	H_a	Type II Error	

The probability of rejecting H_0 computed using the reference distribution is usually called the **Type I error rate** for testing. This is often called "α" and the criterion by which the decision is made is chosen to guarantee that α is small. This makes the standard decision-making methodology asymmetric, more or less "giving H_0 the benefit of any doubt" by requiring that evidence (summarized in the test statistic) for H_a be quite strong before adopting the alternative hypothesis. One chooses α (in advance of testing) to be small (values like $\alpha = .05$ or $\alpha = .01$ are frequently used) and runs only an α probability of rejecting the null hypothesis when it is in fact exactly correct.

The connection between the significance testing and decision-making approaches to hypothesis testing is that to get a test with Type I error rate α, one employs the decision rule

$$\text{"reject } \mathrm{H}_0 \text{ in favor of } \mathrm{H}_\mathrm{a} \text{ if } \quad p\text{-value} < \alpha\text{"}$$

(Again, one rejects the null hypothesis when the sample evidence against it is strong.)

V&J Section 6.2 details a number of critiques of the hypothesis testing paradigm. To simply list some of these to close this section:

1. p-values are highly sample-size-dependent and give no idea of "how wrong" a null hypothesis is,

2. statistical significance is not at all "practical importance" and this fact is often forgotten, and

3. confidence limits implicitly provide testing information *and much more besides.*

13 Small Sample Inference for a Normal Mean μ

(Text Reference/Reading: V&J Section 6.3.1)

The previous two sections of this outline introduced confidence intervals and hypothesis testing, using the case of inference for a single mean *based on a large sample.* A natural next question would be "What can be done if n is not large?" An important answer is based on a probability fact about the random variable

$$T = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{8}$$

when $x_1, x_2, \ldots, x_n$ are iid *from a normal distribution* (with mean μ and standard deviation σ).

While the presence of s (and not σ) in formula (8) prevents the conclusion that T is normal (unless n is large, in which case there is an approximately normal distribution), there *is* an exact known form for the distribution, called the "**Student *t*** distribution." ("Student" was the pen name of the person who first derived the form of the pdf.)

The so-called "t distribution with degrees of freedom parameter ν" has pdf

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\sigma}}\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

The t pdfs are bell-shaped and centered at 0 like the standard normal pdf, but are "flatter"/more spread out than the standard normal. As ν increases they approach the standard normal density as a limit (and for ν of at least 30 or so, their probability assignments are little different from those of the standard normal distribution). Tables and computer functions provide t distribution probabilities and distribution percentage points.

As it turns out, the quantity (8) has (for an underlying normal distribution being sampled) the t distribution with $\nu = n - 1$ degrees of freedom. That means that for a given sample size and desired probability, γ, one may use a table of t distribution percentage points (or a computer routine to evaluate such) to find a number t such that

$$P\left[\text{a } t_{n-1} \text{ random variable is } t \text{ or less}\right] = \gamma$$

This provides confidence limits for a normal mean

$$\bar{x} \pm t\frac{s}{\sqrt{n}}$$

exactly parallel to the large sample limits (6) (based on z rather than t), and t_{n-1} distribution-based p-values for testing H$_0$:$\mu = \#$ (for a normal mean) based on the test statistic

$$T = \frac{\bar{x} - \#}{\frac{s}{\sqrt{n}}}$$

While strictly speaking, these methods provide guaranteed "exact" confidence levels and p-values only when sampling from "exactly normal" distributions, they are generally believed to be fairly "robust." That is, if a distribution/population sampled is not "terribly/ridiculously non-normal," actual/real confidence levels and p-values made using the t_{n-1} distribution percentage points are typically not radically different from the nominal ones (the ones corresponding to normal underlying distributions).

14 Prediction and Tolerance Intervals for a Normal Distribution

(Text Reference/Reading: V&J Section 6.6)

The inference methods for μ of the previous three sections concern using a sample to make properly hedged statements about a distribution *parameter* (its *mean*). A fundamentally different problem (that is nevertheless often confused with the former) is that of using a sample to make properly hedged statements about *likely values of additional observations* (individual measurements/values) from a distribution. We will here briefly consider two versions of this problem:

1. the making of **prediction intervals** (intended to bracket a single additional value from the distribution), and
2. the making of **tolerance intervals** (intended to bracket most of the underlying distribution)

based on samples *from normal distributions.* (These are definitely normal distribution methods, not possessing the kind of "robustness" just mentioned for the t methods of inference for μ. Investigation of the plausibility of a normal distribution underlying a dataset can be approached through normal plotting as covered in Sections 3.2.3 and 5.3 of V&J and briefly reviewed in Section 19 of this outline.)

When sampling from a normal distribution, if $\bar{x}$ and s are based on a sample of size n and a single additional observation x_{new} is drawn from the distribution, it is possible to prove that the random variable

$$\frac{\bar{x} - x_{\text{new}}}{s\sqrt{1 + \dfrac{1}{n}}}$$

has a t_{n-1} distribution. That in turn implies that one or both of the limits

$$\bar{x} \pm ts\sqrt{1 + \frac{1}{n}}$$

can be used to make intervals with a desired confidence for *predicting* x_{new}. (Note that as compared to the t confidence limits for μ, these limits have "an extra 1" under the square root and are much "looser" than the confidence limits for the mean.)

Again when sampling from a normal distribution, if $\bar{x}$ and s are based on a sample of size n, it is possible to derive constants τ_2 and τ_1 (specific to the sample size) so that the two-sided interval with endpoints

$$\bar{x} \pm \tau_2 s$$

and the one-sided intervals

$$(-\infty, \bar{x} + \tau_1 s) \quad \text{and} \quad (\bar{x} - \tau_1 s, \infty)$$

each have a stated confidence of capturing a desired fraction of the underlying *normal* distribution. (Again, these are definitely normal distribution methods, not robust against deviations from normality of the data-generating mechanism.) For example, one can find tabled values τ_2 or τ_1 intended to give 95% confidence in bracketing 99% of the normal distribution that produced the n observations in hand.

15 Inference for a Mean Difference μ_d and for a Difference in Means $\mu_1 - \mu_2$

(Text Reference/Reading: V&J Sections 6.3.2, 6.3.3, 6.3.4)

Two problems often confused by students are those of inference for a *mean difference* and inference for a *difference in means*. In the first case, a single sample of data *pairs* (for example, "before" and "after" or on "treated" and "untreated" versions or on "aspect 1" and "aspect 2" of the same object) can be reduced to differences by subtraction. In the second case, two different samples of single measurements (potentially of different sizes, n_1 and n_2) are gathered with the object of comparison of two corresponding distribution/population means, μ_1 and μ_2.

In the case of inference for a mean difference, data pairs $1, 2, \ldots, n$ are first reduced to differences $d_1, d_2, \ldots, d_n$ that can themselves be processed to make a sample mean, $\bar{d}$, and a sample standard deviation, s_d. Then, the methods of Sections 11 through 13 can be applied to make confidence limits for the mean difference μ_d or to do significance testing for $\mathrm{H}_0{:}\mu_d = \#$. For example, as per Section 13, confidence limits for μ_d are

$$\bar{d} \pm t \frac{s_d}{\sqrt{n}}$$

(Actually, for that matter, the prediction or tolerance limits of the previous section can also be used if there is interest in locating a single additional *difference*, d_{new}, or most of the distribution of d's.)

The case of comparing two means is not simply an application of things that have gone before. Assuming that one has (independent) samples from two different populations (with respective means μ_1 and μ_2) of respective sizes n_1 and n_2, what can be done for inference concerning $\mu_1 - \mu_2$ depends upon the sample sizes. If both are big, then approximate confidence limits for $\mu_1 - \mu_2$ are

$$\bar{x}_1 - \bar{x}_2 \pm z\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

and the hypothesis $\mathrm{H}_0{:}\mu_1 - \mu_2 = \#$ can be tested using the statistic

$$Z = \frac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with approximate p-values obtained from the standard normal distribution.

Where at least one of the sample sizes is small, in Stat 587 we will use methods based on the so-called "Satterthwaite approximation." This treats the variable

$$T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

(which in fact does not have a simple named probability distribution) as approximately t distributed with (random) approximate degrees of freedom given on page 383 of V&J. It turns out that the form in display (6.37) of V&J is at least as big as the smaller of $n_1 - 1$ and $n_2 - 1$. So (as a conservative simplification of what is in second part of Section 6.3.4 of V&J) with

$$\hat{\nu} = \min\left(n_1 - 1, n_2 - 1\right)$$

the limits

$$\bar{x}_1 - \bar{x}_2 \pm t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

(where t is based on $\hat{\nu}$ degrees of freedom) serve as approximate confidence limits for $\mu_1 - \mu_2$, and the statistic

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

can be used to test the hypothesis $\mathrm{H}_0{:}\mu_1 - \mu_2 = \#$ with approximate p-values derived from the t distribution with $\hat{\nu}$ degrees of freedom.

There is a second method of analysis for the small sample version of this problem treated in Section 6.3.4 based on the *additional* assumption that $\sigma_1 = \sigma_2$. It is actually related to analyses we will use for comparison of not just 2, but rather r different means (treated in Chapter 7 of V&J and beginning in Section 20 of this outline). For the case of 2 means, we will use the above formulas, as they are more generally applicable than the ones based on the additional (equal standard deviations) assumption.

16 Inference for Normal Standard Deviations σ, or Variances σ^2

(Text Reference/Reading: V&J Section 6.4)

Sometimes, assessment of the spread of a distribution is more important than assessing the location/center of the distribution. It is thus important to have inference methods for standard deviations. Relatively simple methods are available for data-generating mechanisms that produce normal observations. One- and two-sample versions of these are the subjects of this section.

When sampling from a normal distribution, the (non-negative) quantity

$$\frac{(n-1)\,s^2}{\sigma^2}$$

has a simple probability distribution, for which tables and numerical tools for evaluating probabilities are easy to find. The distribution is called the "chi squared distribution with $\nu = n-1$ degrees of freedom." The χ^2_ν pdf (that is used to produce probabilities) is of the form

$$f(x) = \begin{cases} \dfrac{1}{2^{\nu/2}\Gamma\left(\frac{\nu}{2}\right)} x^{(\nu/2)-1} \exp\left(-\dfrac{x}{2}\right) & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The fact that when sampling a normal distribution $(n-1)\,s^2/\sigma^2 \sim \chi^2_{n-1}$ implies that for L and U respectively small lower and upper percentage points of the χ^2_{n-1} distribution, one or both of the endpoints

$$s\sqrt{\frac{n-1}{U}} \quad \text{and} \quad s\sqrt{\frac{n-1}{L}}$$

can serve as confidence limits for σ. (Of course, confidence limits for variances follow from squaring the values above.) Further, the statistic

$$X^2 = \frac{(n-1)\,s^2}{\#}$$

can be used to test H$_0$:$\sigma^2 = \#$ with p-values derived from the χ^2_{n-1} distribution.

Comparison of two normal distribution standard deviations can be based on the fact that when sampling independently from two normal distributions, the (non-negative) quantity

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has a simple probability distribution, for which tables and numerical tools for evaluating probabilities are easy to find. The distribution is called the "(Snedecor) F distribution with $\nu_1 = n_1 - 1$ (numerator) and $\nu_2 = n_2 - 1$ (denominator) degrees of freedom." The F_{ν_1,ν_2} distribution has pdf

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{(\nu_1/2)-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1 x}{\nu_2}\right)^{(\nu_1+\nu_2)/2}} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

The fact that when sampling independently from two normal distributions $\left(s_1^2/\sigma_1^2\right)/\left(s_2^2/\sigma_2^2\right) \sim F_{n_1-1,n_2-1}$ implies that for L and U respectively small lower and upper percentage points for the F_{n_1-1,n_2-1} distribution, one or both of the endpoints

$$\frac{s_1^2}{U \cdot s_2^2} \text{ and } \frac{s_1^2}{L \cdot s_2^2}$$

can serve as confidence limits for σ_1^2/σ_2^2. (Confidence limits for ratios of standard deviations follow by taking square roots of the values above.) Further, the statistic

$$F = \frac{s_1^2/s_2^2}{\#}$$

can be used to test H$_0$:$\sigma_1^2/\sigma_2^2 = \#$ with p-values derived from the F_{n_1-1,n_2-1} distribution.

Using standard F tables (that provide only upper percentage points) requires knowing that lower percentage points of the F_{n_1-1,n_2-1} distribution can obtained as reciprocals of corresponding upper percentage points of the F_{n_2-1,n_1-1} distribution (the F distribution with numerator and denominator degrees of freedom switched). It is also important to know that two-sided p-values for H$_0$:$\sigma_1^2/\sigma_2^2 = \#$ are usually made by doubling the upper F tail area for the ratio of sample variances made with the larger sample variance in the numerator.

It should be said here that these methods are really only reliable where the underlying distribution is reasonably normal. Again, the normal probability plotting covered in Sections 3.2.3 and 5.3 of V&J and briefly reviewed in Section 19 of this outline is relevant in assessing the plausibility of this circumstance.

17 Inference for Proportions/Binomial Success Probabilities p

(Text Reference/Reading: V&J Section 6.5)

Another pair of important problems of elementary inference are those for a single p and for the difference between two p's $(p_1 - p_2)$. These arise when practical interest centers on fractions of large populations having a particular characteristic or the fractions of outcomes generated by physically stable processes that are of a particular type.

The basic fact that enables inference for a single p is that for $X \sim \text{Bi}(n,p)$, if n is large then X is approximately normal and indeed

$$Z = \frac{X - np}{\sqrt{np\,(1-p)}} = \frac{\frac{X}{n} - p}{\sqrt{\frac{p\,(1-p)}{n}}}$$

is approximately standard normal. Writing

$$\hat{p} = \frac{X}{n}$$

for the sample fraction of "success" outcomes in n trials, this fact leads directly to the (unusable) large sample confidence limits for p,

$$\hat{p} \pm z\sqrt{\frac{p\,(1-p)}{n}}$$

V&J make usable versions of these limits by replacing $p\,(1-p)$ with $\hat{p}\,(1-\hat{p})$. As it turns out, this substitution produces intervals that can for extreme values of p fail to deliver upon nominal confidence levels, the intervals basically tending to be too short.

A modification to this line of reasoning is to replace $\hat{p}\,(1-\hat{p})$ with something a bit larger. A simple choice that works remarkably well is to define

$$\tilde{p} = \frac{n\hat{p} + 2}{n+4} = \frac{X+2}{n+4}$$

(a "sample faction" where 2 fictitious "S" outcomes and 2 fictitious "F" outcomes have been added to the count of X actual successes in n actual trials) and to replace $p\,(1-p)$ with $\tilde{p}\,(1-\tilde{p})$. This leads to large sample confidence limits for p,

$$\hat{p} \pm z\sqrt{\frac{\tilde{p}\,(1-\tilde{p})}{n}}$$

(not presented in V&J, but the best simple formula now known).

Large sample testing for a single p can be done exactly as presented in V&J. The hypothesis H_0:$p = \#$ can be tested using the test statistic

$$Z = \frac{\hat{p} - \#}{\sqrt{\frac{\#\,(1-\#)}{n}}}$$

and approximate p-values derived from the standard normal distribution.

Two large (independent) samples from populations or processes with underlying proportions p_1 and p_2 producing sample proportions of successes (among respectively n_1 and n_2 trials) $\hat{p}_1$ and $\hat{p}_2$ can be used to do inference for $p_1 - p_2$. The same logic that enables inference for a single p produces large sample confidence limits for $p_1 - p_2$,

$$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\frac{\tilde{p}_1\left(1-\tilde{p}_1\right)}{n_1} + \frac{\tilde{p}_2\left(1-\tilde{p}_2\right)}{n_2}}$$

(where the $\tilde{p}$'s are as for the single sample case). And hypothesis testing for H_0:$p_1 - p_2 = 0$ can be done exactly as presented in V&J. With

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

the test statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\left(1-\hat{p}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

can be used with p-values derived from an approximately standard normal reference distribution.

While some small sample formulas methods exist for these inference problems, they are not particularly simple. More importantly, they are not typically of practical importance. Small numbers of S/F outcomes provide very little information about underlying p's, and inferences based on them rarely provide definitive practical conclusions.

18 One- and Two-Sample Inference Formula Summary

Inference For	Sample Size	Assumptions	H_0, Test Stat, Reference	Interval	Section
μ (one mean)	large n		$H_0{:}\mu = \#$ $Z = \dfrac{\bar{x} - \#}{s/\sqrt{n}}$ standard normal	$\bar{x} \pm z\dfrac{s}{\sqrt{n}}$	6.1, 6.2
	small n	observations normal	$H_0{:}\mu = \#$ $T = \dfrac{\bar{x} - \#}{s/\sqrt{n}}$ t with $\nu = n - 1$	$\bar{x} \pm t\dfrac{s}{\sqrt{n}}$	6.3
$\mu_1 - \mu_2$ (difference in means)	large n_1, n_2	independent samples	$H_0{:}\mu_1 - \mu_2 = \#$ $Z = \dfrac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ standard normal	$\bar{x}_1 - \bar{x}_2 \pm z\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$	6.3
	small n_1 or n_2	independent normal samples	$H_0{:}\mu_1 - \mu_2 = \#$ $T = \dfrac{\bar{x}_1 - \bar{x}_2 - \#}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ t with $\hat{\nu}$ given on page 383, or just $\hat{\nu} = \min(n_1 - 1, n_2 - 1)$	$\bar{x}_1 - \bar{x}_2 \pm \hat{t}\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$ use random $\hat{\nu}$ given on page 383, or just $\hat{\nu} = \min(n_1 - 1, n_2 - 1)$	6.3
μ_d (mean difference)	large n	(paired data)	$H_0{:}\mu_d = \#$ $Z = \dfrac{\bar{d} - \#}{s_d/\sqrt{n}}$ standard normal	$\bar{d} \pm z\dfrac{s_d}{\sqrt{n}}$	6.3
	small n	(paired data) normal differences	$H_0{:}\mu_d = \#$ $T = \dfrac{\bar{d} - \#}{s_d/\sqrt{n}}$ t with $\nu = n - 1$	$\bar{d} \pm t\dfrac{s_d}{\sqrt{n}}$	6.3

Inference For	Assumptions	Interval	Section
x_{new} (a single additional value)	observations normal	$\bar{x} \pm ts\sqrt{1 + \frac{1}{n}}$	6.6
most of the distribution	observations normal	$\bar{x} \pm \tau_2 s$ or $(\bar{x} - \tau_1 s, \infty)$ or $(-\infty, \bar{x} + \tau_1 s)$	6.6

Inference For	Assumptions	H$_0$, Test Stat, Reference	Interval	Section
σ^2 (one variance)	observations normal	$\mathrm{H}_0\colon \sigma^2 = \#$ $X^2 = \dfrac{(n-1)s^2}{\#}$ χ^2 with $\nu = n-1$	$\dfrac{(n-1)s^2}{\chi^2_{\text{upper}}}$ and/or $\dfrac{(n-1)s^2}{\chi^2_{\text{lower}}}$	6.4
σ_1^2/σ_2^2 (variance ratio)	observations normal independent samples	$\mathrm{H}_0\colon \dfrac{\sigma_1^2}{\sigma_2^2} = \#$ $F = \dfrac{s_1^2/s_2^2}{\#}$ F with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$	$\dfrac{s_1^2}{F_{\text{upper}} \cdot s_2^2}$ and/or $\dfrac{s_1^2}{F_{\text{lower}} \cdot s_2^2}$	6.4

Inference For	Sample Size/Assumptions	H$_0$, Test Stat, Reference	Interval	Section
p (one proportion)	large n	$\mathrm{H}_0\colon p = \#$ $Z = \dfrac{\hat{p} - \#}{\sqrt{\dfrac{\#(1-\#)}{n}}}$ standard normal	$\hat{p} \pm z\sqrt{\dfrac{\tilde{p}(1-\tilde{p})}{n}}$ use $\tilde{p} = \dfrac{n\hat{p}+2}{n+4}$	6.5
$p_1 - p_2$ (difference in proportions)	large n_1, n_2 independent samples	$\mathrm{H}_0\colon p_1 - p_2 = 0$ $Z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})}\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ use $\hat{p}$ given in display (6.71) page 411 standard normal	$\hat{p}_1 - \hat{p}_2 \pm z\sqrt{\dfrac{\tilde{p}_1(1-\tilde{p}_1)}{n_1} + \dfrac{\tilde{p}_2(1-\tilde{p}_2)}{n_2}}$ use $\tilde{p}_1 = \dfrac{n_1\hat{p}_1+2}{n_1+4}$ and $\tilde{p}_2 = \dfrac{n_2\hat{p}_2+2}{n_2+4}$	6.5

19 Q-Q Plotting and Probability Plotting (e.g. Normal Plotting)

(Text Reference/Reading: V&J Sections 3.2.3, 5.3)

The comparison of "shapes" for distributions is a basic statistical activity. The *most* important version of this is comparison of the shape of an empirical distribution (the "shape" of a dataset) to the shape of a theoretical/probability distribution. One expects the shape of a dataset to be indicative of the nature of a corresponding underlying data-generating mechanism. So if (for example) one intends to use statistical methodology built on a mathematical assumption of normality, checking to see that a sample is not terribly non-normal-looking is exercise of due diligence in attempting to not make unjustified conclusions.

The comparison of shapes of two datasets (two empirical distributions) is less important than comparison of an empirical shape to a theoretical shape, but the same methodology is used to do both (and this methodology easiest to understand for the empirical versus empirical case). It is based on the notion of a distribution "quantile" $Q(p)$. In rough terms, this is a number that places a fraction p of the distribution to the left and a fraction $1-p$ of the distribution to the right. The exact convention used in Stat 587 to define quantiles of a finite dataset is discussed in V&J. (For p of the form $(i-.5)/n$ for integer i and sample size n, $Q(p)$ is the ith smallest value in the sample. Other quantiles are defined by linear interpolation.)

A Q-Q plot is then a plot of ordered pairs

$$(Q_1(p), Q_2(p))$$

for some appropriate set of values of p. For the case of two datasets, the values of $p=(i-.5)/n$ for n the smaller of the two sample sizes are typically used. For the case of an empirical distribution and a theoretical one, the values $p=(i-.5)/n$ are used.

In the important special case where assessing agreement with the normal distributional shape is in view, standard normal quantiles $Q_z(p) = \Phi^{-1}(p)$ are employed for the vertical plotting positions. Supposing that x_i is the ith smallest ordered data value, the n points plotted on a **normal (probability) plot** are then of the form

$$\left(x_i, Q_z\left(\frac{i-.5}{n}\right)\right)$$

What makes a Q-Q plot informative is the fact that *equality of shape for two distributions is equivalent to them having linearly related quantile functions.* So linearity on a Q-Q plot is indicative of equality of shape for the two distributions being considered. And departures from linearity are potentially interpretable as highlighting various kinds of differences in shape (such as a "long tail" of one distribution relative to the other).

20 The One-Way Normal Model, Residuals, and Pooled Sample Standard Deviation s_{P}

(Text Reference/Reading: V&J Section 7.1) (Also V&J SMQA Section 5.1.1)

The ultimate goal in Stat 587 is the consideration of data from *multifactor* studies where the object is quantification of the impact of those variables on some response, y. In those contexts, every different set of values for the (multiple) factors defines a different "sample" of y's. That is, practical multifactor studies are of necessity *multisample* studies. Before digging into the specifics of different kinds of multifactor statistical methodologies, we begin here by considering what can be said in general about inference based on r-sample (for $r > 1$) studies *without reference to any specific pattern or structure associated with factors defining the multiple samples.* Chapter 7 of V&J terms this material the analysis of "unstructured" multisample studies. Somewhat more common terminology refers to these basic methods as "one-way" methods.

The most commonly used statistical model for r-sample data is the "one-way normal model." In words, that model says that each of r different (sets of) conditions independently produces normally distributed observations with means that may differ, but whose standard deviations are all the same. In symbols, if

$$y_{ij} = \text{the } j\text{th observation from (set of) condition(s) } i$$

then for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, n_i$ for each i, the observations are independent with

$$y_{ij} \sim \mathrm{N}\left(\mu_i, \sigma^2\right)$$

for parameters $\mu_1, \mu_2, \ldots, \mu_r$ and σ^2. Or if ϵ_{ij} for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, n_i$ for each i are iid $\mathrm{N}\big(0, \sigma^2\big)$ random errors,

$$y_{ij} = \mu_i + \epsilon_{ij} \tag{9}$$

Representation (9) is intuitively attractive, in that it partitions what is observed (y_{ij}) into a kind of "signal" (μ_i) plus "noise" (ϵ_{ij}). The magnitude of the noise is governed by the parameter σ, and typically the main goal of statistical analysis is understanding any interpretable patterns extant in the signal.

Where sample sizes $n_1, n_2, \ldots, n_r$ are all not small, a way of investigating the plausibility of the basic one-way normal model is to make normal plots of the r different samples on a single set of axes (looking for more-or-less parallel more-or-less straight-line plots). But often, samples sizes in multisample studies are small and something else must be done to make sanity checks on the model assumptions. What is standard is based on so-called "residuals." That is, for

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

the sample mean of observations from the ith (set of) condition(s), the **residual**

$$e_{ij} = y_{ij} - \bar{y}_i$$

is an approximation for the error

$$\epsilon_{ij} = y_{ij} - \mu_i$$

(The residual is what is "left over" in an observation after accounting for the apparent/approximate signal $\bar{y}_i$.)

Since the model (9) says that the errors ϵ_{ij} are iid $\mathrm{N}\left(0,\sigma^2\right)$ random variables, one can expect their approximations, the residuals e_{ij}, to look more or less like a random sample from a normal distribution with mean 0. So various kinds of **plotting of residuals** is done, hoping to see plots consistent with this expectation. A linear normal plot of residuals and lack of any obvious pattern or trend on plots of residuals against variables not of interest (like time order of observation, or some extraneous experimental condition like ambient temperature, etc.) is what one hopes to find.

As we've already said, the parameter σ, governs the level of "noise" through which important changes in mean response must be seen. It is then essential to estimate this parameter. Under the one-way model assumptions, every sample standard deviation serves to estimate the same σ. As such, it makes sense to "pool" the information they all carry into a single estimate of σ. To this end, we define the pooled sample variance (a weighted average of the individual sample variances)

$$\begin{aligned} s_{\mathrm{P}}^2 &= \frac{(n_1-1)\,s_1^2+(n_2-1)\,s_2^2+\cdots+(n_r-1)\,s_r^2}{(n_1-1)+(n_2-1)+\cdots+(n_r-1)} \\ &= \frac{\sum_{i,j}\left(y_{ij}-\bar{y}_i\right)^2}{n-r} \end{aligned}$$

(using the notation n for the total number of observations, $\sum_{i=1}^{r} n_i$, in the last line here). Taking the square root, we get a pooled sample standard deviation

$$s_{\mathrm{P}} = \sqrt{s_{\mathrm{P}}^2}$$

(sometimes called a "root mean squared error").

Under the one-way normal model assumptions, it's possible to make confidence limits for σ of the form

$$s_{\mathrm{P}}\sqrt{\frac{n-r}{U}} \quad \text{and/or} \quad s_{\mathrm{P}}\sqrt{\frac{n-r}{L}}$$

for U and/or L small (upper and/or lower) χ^2_{n-r} percentage points. These provide bounds on the level of "background noise" in a multisample study.

A slight refinement of the development of residuals concerns the fact that while the unobservable errors ϵ_{ij} have the same variance (σ^2), the residuals e_{ij} do not have the same variance. It turns out that $\mathrm{Var}e_{ij} = \dfrac{n_i-1}{n_i}\sigma^2$, which potentially varies with i. So, sometimes, instead of plotting with ordinary residuals e_{ij} one removes this issue (by standardization) and plots with **standardized residuals**

$$e_{ij}^* = \frac{e_{ij}}{s_{\mathrm{P}}\sqrt{\dfrac{n_i-1}{n_i}}}$$

hoping to see approximately "*standard*-normal-looking" plots.

21 Confidence Intervals for Linear Combinations of Means

(Text Reference/Reading: V&J Section 7.2) (See also V&J SMQA Section 5.1.2)

The pooled sample standard deviation identifies the level of background noise against which observed differences in sample means in an r-sample study are to be judged. This section provides a basic technical tool for making such judgments. That is the making of confidence limits for a linear combination of condition means.

For any set of constants $c_1, c_2, \ldots, c_r$ define the linear combination of model means

$$L = c_1\mu_1 + c_2\mu_2 + \cdots + c_r\mu_r \tag{10}$$

A natural data-based approximation to this is the corresponding linear combination of sample means

$$\hat{L} = c_1\bar{y}_1 + c_2\bar{y}_2 + \cdots + c_r\bar{y}_r \tag{11}$$

Then, under the one-way normal model, one or both of the values

$$\hat{L} \pm t s_{\mathrm{P}} \sqrt{\sum_{i=1}^{r} \frac{c_i^2}{n_i}} \tag{12}$$

for t a small upper percentage point of the t distribution with $\nu = n - r$ degrees of freedom can be used as confidence limits for L. (Rationale for this formula is on pages 464 and 465 of V&J.)

Particular special cases of this development provide important simple intervals for a multisample study. Where a single c_i is 1 and all others are 0, one has confidence limits for a single mean. Where one c_i is 1, another is -1, and all others are 0, one has confidence limits for a difference in two particular means.

22 One-Way ANOVA

(Text Reference/Reading: V&J Section 7.4.1-7.4.3)

V&J briefly discusses testing the hypothesis $\mathrm{H}_0{:}L = \#$ for a particular set of values $c_1, c_2, \ldots, c_r$. But by far the most commonly considered hypothesis testing problem in the one-way model is that for $\mathrm{H}_0{:}\mu_1 = \mu_2 = \cdots = \mu_r$, namely that there are no differences at all among the distributions of responses for the r conditions studied. There is a standard F test for this problem that additionally provides important intuition about "kinds of variation" seen among the n responses y_{ij}.

Consider first the overall/grand sample mean computed ignoring the sample boundaries

$$\bar{y} = \frac{1}{n}\sum_{i,j} y_{ij}$$

(Note that this is not in general the same as the arithmetic average of the $\bar{y}_i$'s.) As a measure of the variation among the $\bar{y}_i$, we then take a sum of squared deviations of these from the grand mean

$$\sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2$$

This is a possible quantification of "between-sample variation." It is big when there are large differences among the sample means, indicating large "signals" in the data.

Ultimately, to test $\mathrm{H}_0{:}\mu_1 = \mu_2 = \cdots = \mu_r$ one can use the test statistic

$$\begin{aligned} F &= \frac{\dfrac{1}{r-1}\displaystyle\sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2}{s_{\mathrm{P}}^2} \\ &= \frac{\dfrac{1}{r-1}\displaystyle\sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2}{\dfrac{1}{n-r}\displaystyle\sum_{i,j} \left(y_{ij} - \bar{y}_i\right)^2} \end{aligned} \tag{13}$$

and an $F_{r-1,n-r}$ reference distribution, where p-values are upper tail areas beyond the sample value of F produced by the data in hand.

The sums in the numerator and denominator of the F statistic (13) have an illuminating relationship to an overall/grand sample variance (computed ignoring the sample boundaries). That is, for s^2 the overall sample variance, it is an algebraic fact that

$$(n-1)\, s^2 = \sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2 + (n-r)\, s_{\mathrm{P}}^2$$

In other symbols, this is

$$\sum_{i,j} \left(y_{ij} - \bar{y}\right)^2 = \sum_{i=1}^{r} n_i \left(\bar{y}_i - \bar{y}\right)^2 + \sum_{i,j} \left(y_{ij} - \bar{y}_i\right)^2$$

These are versions of the so-called "one-way ANOVA identity." (ANOVA is standard jargon for "ANalysis Of VAriance.") The terms in this identity are called "sums of squares." The first is called a "total" sum of squares. The second is usually called a "treatment" sum of squares. The third is called the "error" sum of squares. In this language, the identity is

$$SSTot = SSTr + SSE$$

As a way of both organizing the computation of the F statistic (13) *and* providing additional intuition about partitioning of both observed variation in response and "degrees of freedom," it is common to summarize the computation of the one-way ANOVA F statistic in a so-called "ANOVA Table." (There are actually many possible ANOVA tables in applied statistics. The one appropriate here is in Table 3.) The "MS" column (that has sums of squares divided by degrees of freedom in it) is a "mean square" column.

Table 3: General Form of the One-Way ANOVA Table

ANOVA Table for Testing H$_0$:$\mu_1 = \mu_2 = \cdots = \mu_r$

Source	SS	df	MS	F
Treatments	$SSTr$	$r-1$	$SStr/\left(r-1\right)$	$MSTr/MSE$
Error	SSE	$n-r$	$SSE/\left(n-r\right)$	
Total	$SSTot$	$n-1$		

Part II

Classical Multifactor Data Analysis: Regression and Factorial Analyses

23 Simple Linear Regression (SLR) Introduction- Least Squares, the Sample Correlation, R^2, and Residuals

(Text Reference/Reading: V&J Section 4.1)

We now begin to consider quantifying how a mean response changes with values of one or more explanatory variables. We start with the simplest possible case, that where there is a *single quantitative factor/variable* (call it x) and the relationship between x and the response y is approximately *linear*. To be more explicit, we assume that n data pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ provide information on an approximate relationship

$$y \approx \beta_0 + \beta_1 x$$

(Plotting of the n data pairs is the obvious place to begin data analysis, determining that approximate linearity is an appropriate description of the relationship.)

The classical method of using the n data pairs to choose a slope and intercept to represent a relationship between x and y is to employ the **least squares criterion**. This means choosing β_0 and β_1 to minimize the quadratic function of two variables

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Provided the x_i are not all the same, setting partial derivatives of $S(\beta_0, \beta_1)$ equal to 0 and solving for β_0 and β_1 produces the "least squares coefficients"

$$b_1 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(the slope) and

$$b_0 = \bar{y} - b_1 \bar{x}$$

(the intercept). Then (using $\hat{y}$ to stand for the fitted or predicted response) the equation of the least squares line is

$$\hat{y} = b_0 + b_1 x$$

and we write

$$\hat{y}_i = b_0 + b_1 x_i$$

for the value of fitted or predicted y for the ith data case.

It is useful to have measures of how well a fitted line does at describing the n data pairs (x_i, y_i). One such measure is the sample correlation between x and y. This is

$$r = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

As it turns out, $-1 \leq r \leq 1$ and $|r| = 1$ exactly when all plotted points (x_i, y_i) fall on a single straight line. The case $r = 1$ is the case where the line has a positive slope and the the case $r = -1$ is the case where the line has a negative slope.

Another measure of strength of apparent linear relationship is the so-called "coefficient of determination." This is

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 - \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

that has an interpretation as "the fraction of the raw variation in y accounted for using the (linear in x) prediction equation." This can be expressed in other notation using the sums of squares

$$\begin{aligned} SSTot &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\ SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{, and} \\ SSR &= SSTot - SSE \end{aligned} \tag{14}$$

and then

$$R^2 = \frac{SSR}{SSTot}$$

As it turns out, R^2 also has an interpretation as a squared sample correlation. It is the squared correlation between y and $\hat{y}$. (Further, since $\hat{y}$ and x are perfectly correlated, R^2 is also the squared sample correlation between x and y in this simple case. This last interpretation is one that is special to the present single-explanatory-variable case.)

There is also a notion of "residuals" for the least squares fitting of the line. That is, the residuals in this context are values

$$e_i = y_i - \hat{y}_i \tag{15}$$

(the sum of whose squares make up SSE). These can be plotted and interpreted in the same kinds of ways that were indicated in Section 20 of this outline for residuals in the one-way model.

24 The Normal Simple Linear Regression Model and Inference for σ

(Text Reference/Reading: V&J Section 9.1.1)

The one-way normal model (9) imposes no restrictions on the means μ_i. Inference for approximately linear relationships between x and y employ a specialization of the basic "independent normal observations with a common variance" assumptions of Section 20. That is, we now adopt the assumption that the mean value for y is linear in x. This can be written as

$$\mu_{y|x} = \beta_0 + \beta_1 x \tag{16}$$

A complete specification of the "normal simple linear regression model" is then that for $i = 1, 2, \ldots, n$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{17}$$

for iid $\mathrm{N}(0, \sigma^2)$ random "errors" ϵ_i. (Model statement (17) is, of course, exactly parallel to the less restrictive statement (9). The present model has 3 parameters, β_0, β_1, and σ. The earlier one had $r+1$ parameters, $\mu_1, \mu_2, \ldots, \mu_r$, and σ.)

An estimate of σ can be built from SSE. That is, a line-fitting sample variance is

$$s_{\mathrm{LF}}^2 = \frac{1}{n-2} SSE = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and the corresponding line-fitting sample standard deviation is

$$s_{\mathrm{LF}} = \sqrt{s_{\mathrm{LF}}^2}$$

This can, in turn, be used to make confidence limits for σ in the normal simple linear regression model. Under model (17), values

$$s_{\mathrm{LF}} \sqrt{\frac{n-2}{U}} \quad \text{and/or} \quad s_{\mathrm{LF}} \sqrt{\frac{n-2}{L}}$$

for U and/or L small (upper and/or lower) χ^2_{n-2} percentage points serve as confidence limits for σ.

25 Inference for the SLR Slope β_1 and Mean y at a Given x, Prediction of $y_{\mathbf{new}}$ at x, and Standardized Residuals

(Text Reference/Reading: V&J Sections 9.1.2-9.1.4)

The parameter β_1 in the simple linear regression model (17) represents the rate of change of mean y with respect to x and thus measures the impact that changes in x have on the mean response. Inference for it is an important part of a typical simple linear regression analysis. Confidence intervals for β_1 can be made using one or both of the endpoints

$$b_1 \pm t \frac{s_{\mathrm{LF}}}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}$$

for t a small upper percentage point of the t distribution with $\nu = n-2$ degrees of freedom. Further, the hypothesis H_0:$\beta_1 = \#$ can be tested using the test statistic

$$T = \frac{b_1 - \#}{\dfrac{s_{\mathrm{LF}}}{\sqrt{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}}$$

and a t_{n-2} reference distribution. Note that under the SLR model, if $\beta_1 = 0$, the mean response doesn't change with x. So testing H_0:$\beta_1 = 0$ is a way of addressing the question of whether x has a discernible impact on the value of mean y. (Other common language is that of asking whether "x is of any use in 'explaining' or 'predicting' y.")

The quantity $\mu_{y|x} = \beta_0 + \beta_1 x$ first defined in display (16) is another important object in most SLR analyses. For a given input value x, this is the average "system response." Confidence intervals for it can be made using one or both of the endpoints

$$\hat{y} \pm t s_{\mathrm{LF}} \sqrt{\frac{1}{n} + \frac{\left(x-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}$$

for t a small upper percentage point of the t distribution with $\nu = n-2$ degrees of freedom. Further, the hypothesis H_0:$\mu_{y|x} = \#$ can be tested using the test statistic

$$T = \frac{\hat{y} - \#}{s_{\mathrm{LF}} \sqrt{\dfrac{1}{n} + \dfrac{\left(x-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}}}$$

and a t_{n-2} reference distribution. Note, by the way, that the case of $x = 0$ provides inferences for β_0, the intercept in the SLR model. Further, one intuitively plausible implication of this formula is that most is known (limits for mean y are tightest) at $x = \bar{x}$ (where the final term under the root is 0).

Simple linear regression has its own form of prediction intervals, aiming to capture a new value of the response, y_{new}, for a particular value of the input, x. These, like the one-sample prediction limits of Section 14 of this outline, are related to confidence limits for a mean response by the "addition of 1 under a square root." That is, prediction intervals for y_{new} at x can be made using one or both of the endpoints

$$\hat{y} \pm t s_{\text{LF}} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

for t a small upper percentage point of the t distribution with $\nu = n - 2$ degrees of freedom.

While plotting with SLR residuals $e_i = y_i - \hat{y}_i$ is possible, it is common to correct them for their lack of a common variance and plot instead with standardized residuals

$$e_i^* = \frac{e_i}{s_{\text{LF}} \sqrt{1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}}$$

26 ANOVA and SLR

(Text Reference/Reading: V&J Section 9.1.5)

The definition of R^2 in terms of sums of squares already begins to hint that there is a form of analysis of variance associated with SLR much like that seen in Section 22. As there, we begin with an F test of the hypothesis that all mean responses are the same. In the SLR context, that is the hypothesis $\mathrm{H}_0{:}\beta_1 = 0$. We've already noted that a t test of this is possible. Here we note that with a two-sided alternative hypothesis, one may use a test statistic

$$F = \frac{SSR/1}{SSE/(n-2)}$$

and an $F_{1,n-2}$ reference distribution. (Two-sided p-values are right tail areas beyond the observed value of this statistic.)

Rearranging the definition (14) one has an ANOVA identity appropriate to SLR

$$SSTot = SSR + SSE$$

And the computation of the F statistic can be summarized (and the partitioning of $SSTot$ appropriate to SLR presented) in an ANOVA table. The general version of this for SLR is presented in Table 4.

Table 4: General Form of the ANOVA Table for Simple Linear Regression

ANOVA Table for Testing $\mathrm{H}_0{:}\beta_1 = 0$

Source	SS	df	MS	F
Regression (on x)	SSR	1	$SSR/1$	MSR/MSE
Error	SSE	$n-2$	$SSE/(n-2)$	
Total	$SSTot$	$n-1$		

As it turns out, the F statistic for testing $\mathrm{H}_0{:}\beta_1 = 0$ is the square of the t statistic for the hypothesis, and the p-values produced (for a two-sided alternative hypothesis) are the same.

The organization provided by Table 4 provides intuition about what is being said by $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Large observed F correspond to large SSR, and in turn to large R^2. The table entry $SSE/(n-2)$ is exactly s^2_{LF} and is a "mean squared error." $SSTot$ and degrees of freedom $n-1$ are partitioned according to sources "explained" and "left over."

27 Multiple Linear Regression (MLR) Introduction- Least Squares, R^2, Residuals, the MLR Model, and Inference for σ^2

(Text Reference/Reading: V&J Sections 4.2,9.2.1&9.2.5)

"**Multiple Linear Regression**" is in many ways the natural extension of the simple linear regression material just presented. To begin, the basic data available for analysis are n vectors (of dimension $k+1$)

$$(x_{11}, x_{21}, \ldots, x_{k1}, y_1), (x_{12}, x_{22}, \ldots, x_{k2}, y_2), \ldots, (x_{1n}, x_{2n}, \ldots, x_{kn}, y_n)$$

The approximate relationship between the input/predictor/explanatory variables/(quantitative) factors x_j (for $j = 1, 2, \ldots, k$) employed in this methodology is the linear form

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Using least squares and the n data vectors to find appropriate values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ means choosing β_0 through β_k to minimize the quadratic function of $k+1$ variables

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}))^2$$

The set of $k+1$ equations

$$\frac{\partial}{\partial_l} S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = 0 \quad \text{for } l = 0, 1, 2, \ldots, k$$

is called the set of normal (perpendicular) equations and can typically be solved uniquely for $k+1$ minimizing (least squares) coefficients

$$b_0, b_1, b_2, \ldots, b_k$$

There are no simple formulas for these (unless matrix notation is used, something that will not be done in Stat 587), but it is easy enough to get any decent statistical package to provide these fitted coefficients.

Using the notation

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \tag{18}$$

and in particular

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} \quad ,$$

the MLR version of "SSE" looks just like the SLR version, namely

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad .$$

Then, since

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

has nothing to do with what approximate relationship between y and a set of explanatory factors is under discussion, the obvious regression sum of squares for MLR is just as for SLR,

$$SSR = SSTot - SSE$$

Then (as in SLR) the fraction of raw variability accounted for by the fitted (multiple linear regression) equation (18) is

$$R^2 = \frac{SSR}{SSTot}$$

As in SLR, this also turns out to be a squared sample correlation between y and $\hat{y}$ (but has no interpretation as a squared correlation between y and any individual predictor, x_j).

MLR residuals have the same form as SLR residuals, namely

$$e_i = y_i - \hat{y}_i \tag{19}$$

and their plotting is possible, though typically standardized versions based on a generalization of the normal SLR model are employed.

The **normal multiple linear regression** model is another specialization of the one-way normal model (9). It generalizes the SLR model (17) by allowing the mean response to depend linearly on k explanatory variables (rather than just one). That is, it is built on the assumption that

$$\mu_{y|x_1,x_2,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{20}$$

(Notice that when all but one of $\beta_1, \beta_2, \ldots, \beta_k$ are 0, relationship (20) reduces to the SLR assumption (16).) A complete specification of the "normal multiple linear regression model" is then that for $i = 1, 2, \ldots, n$ and (known/fixed) input vectors $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \epsilon_i \tag{21}$$

for iid $\mathrm{N}(0, \sigma^2)$ random "errors" ϵ_i. (Model statement (21) is, of course, a generalization of the SLR model (17).) The present model has $k+2$ parameters, $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ and σ. β_0 is an intercept. $\beta_1, \beta_2, \ldots, \beta_k$ are rates of change in mean y with respect to a single predictor *with all others held fixed.* The standard deviation, σ, governs how much variation is seen in response *when all predictors are held fixed.*

An estimate of σ can be built from SSE. That is, a surface-fitting sample variance is

$$s^2_{\mathrm{SF}} = \frac{1}{n-(k+1)} SSE = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

and the corresponding surface-fitting sample standard deviation is

$$s_{\mathrm{SF}} = \sqrt{s^2_{\mathrm{SF}}}$$

This can, in turn, be used to make confidence limits for σ in the normal multiple linear regression model. Under model (21), values

$$s_{\mathrm{SF}} \sqrt{\frac{n-k-1}{U}} \quad \text{and/or} \quad s_{\mathrm{SF}} \sqrt{\frac{n-k-1}{L}}$$

for U and/or L small (upper and/or lower) χ^2_{n-k-1} percentage points serve as confidence limits for σ.

28 Inference for the MLR Coefficients β_l and Mean y at a Set of Values $x_1, x_2, \ldots, x_k$, Prediction of $y_{\mathbf{new}}$ at $x_1, x_2, \ldots, x_k$, and Standardized Residuals

(Text Reference/Reading: V&J Sections 9.2.1-9.2.4)

Continuing with material parallel to that presented for SLR, consider estimation of individual regression coefficients β_l. As it turns out, there is a standard error (estimated standard deviation) for b_l that we will call "se_{b_l}" and is a multiple of s_{SF}. (The multiple depends upon the data only through the values of the $(x_{1i}, x_{2i}, \ldots, x_{ki})$ in the dataset.) There is no simple formula for se_{b_l} (unless one is willing to use matrix notation). In particular, **no "by hand" formula from SLR is relevant here**! But MLR programs will compute and print out numerical values for these standard errors and it is possible to argue that under the normal MLR model the random variable

$$T = \frac{b_l - \beta_l}{\mathrm{se}_{b_l}}$$

has a t_{n-k-1} distribution. This in turn implies that one or both of the values

$$b_l \pm t \cdot \mathrm{se}_{b_l}$$

(for t a small upper percentage point of the t_{n-k-1} distribution) can be used to make a confidence interval for β_l. Further, the hypothesis H_0:$\beta_l = \#$ can be tested using the statistic

$$T = \frac{b_l - \#}{\mathrm{se}_{b_l}}$$

and a t_{n-k-1} reference distribution. (The case of $\# = 0$ is most common, as the corresponding hypothesis implies that $\mu_{y|x_1,x_2,\ldots,x_k}$ doesn't depend upon x_l.)

It is also possible to identify a standard error for $\hat{y}$ that is a multiple of s_{SF}. (The multiplier again depends upon the data only through the values of the $(x_{1i}, x_{2i}, \ldots, x_{ki})$ in the dataset.) We will call this "$\mathrm{se}_{\hat{y}}$" and recognize that while there is no simple formula for it, MLR programs will compute it and print it out. It is possible to argue that under the normal MLR model the random variable

$$T = \frac{\hat{y} - \mu_{y|x_1,x_2,\ldots,x_k}}{\mathrm{se}_{\hat{y}}}$$

has a t_{n-k-1} distribution. This in turn implies that one or both of the values

$$\hat{y} \pm t \cdot \mathrm{se}_{\hat{y}}$$

(for t a small upper percentage point of the t_{n-k-1} distribution) can be used to make a confidence interval for $\mu_{y|x_1,x_2,\ldots,x_k}$. Further, the hypothesis H_0:$\mu_{y|x_1,x_2,\ldots,x_k} = \#$ can be tested using the statistic

$$T = \frac{\hat{y} - \#}{\mathrm{se}_{\hat{y}}}$$

and a t_{n-k-1} reference distribution. As for SLR, the choice of $x_1 = 0, x_2 = 0, \ldots, x_k = 0$ provides inference methods for the intercept, β_0.

The standard error for $\hat{y}$ can also be used to produce prediction limits for y_{new} at $x_1, x_2, \ldots, x_k$. It is the case that

$$T = \frac{\hat{y} - y_{\text{new}}}{\sqrt{s_{\text{SF}}^2 + \text{se}_{\hat{y}}^2}}$$

has a t_{n-k-1} distribution. This in turn implies that one or both of the values

$$\hat{y} \pm t\sqrt{s_{\text{SF}}^2 + \text{se}_{\hat{y}}^2}$$

(for t a small upper percentage point of the t_{n-k-1} distribution) can be used to make a prediction interval for y_{new} at $x_1, x_2, \ldots, x_k$.

It also turns out that the standard error for $\hat{y}_i$ is helpful in producing standardized residuals for the normal MLR model. That is, corresponding to the MLR residuals (19) are *standardized* residuals

$$e_i^* = \frac{e_i}{\sqrt{s_{\text{SF}}^2 - \text{se}_{\hat{y}_i}^2}}$$

that correct the residuals by giving them a common variance. Plotting these (expecting approximately-standard-normal behavior if the normal MLR model is appropriate) is an improvement over the plotting of ordinary residuals. Most MLR programs will produce them more or less automatically.

29 MLR and ANOVA-Overall/Full and Partial F Tests

(Text Reference/Reading: V&J Section 9.2.5)

MLR has its ANOVA methodology and corresponding intuition and F tests. There are both an "overall" F test and potentially many "partial" F tests (and associated breakdowns of $SSTot$). We begin with the overall ANOVA and test.

In a manner parallel to that met in Section 26, the hypothesis that all of the "slopes" β_l are 0 is the hypothesis that mean y doesn't change with any of the explanatory variables. That is, in the MLR model, H$_0$:$\beta_1 = \beta_2 = \cdots = \beta_k = 0$ is the hypothesis that mean y is constant at β_0. This hypothesis can be tested using the statistic

$$F = \frac{SSR/k}{SSE/\left(n-k-1\right)}$$

and an $F_{k,(n-k-1)}$ reference distribution. Where SSR is large (and thus, SSE is small), R^2 is large and the fitted linear equation is interpreted as accounting for much of the observed variation in y, so upper tail areas beyond an observed value for F are used as p-values. This test can be thought of as providing an "observed significance level" for R^2, in that

$$F = \frac{R^2/k}{\left(1-R^2\right)/\left(n-k-1\right)}$$

The computation of the overall F statistic can be summarized (and the partitioning of $SSTot$ appropriate to MLR presented) in an ANOVA table. The general version of this for MLR is presented in Table 5.

Table 5: General Form of the ANOVA Table for Mulitple Linear Regression

ANOVA Table for Testing H$_0$:$\beta_1 = \beta_2 = \cdots = \beta_k = 0$

Source	SS	df	MS	F
Regression (on $x_1, x_2, \ldots, x_k$)	SSR	k	SSR/k	MSR/MSE
Error	SSE	$n-k-1$	$SSE/\left(n-k-1\right)$	
Total	$SSTot$	$n-1$		

The table entry $SSE/\left(n-k-1\right)$ is exactly s^2_{SF} and is the "mean squared error" for MLR. $SSTot$ and degrees of freedom $n-1$ are partitioned according to sources "explained" and "left over" in the fitting of the MLR equation.

The overall F test and associated ANOVA concerns the question of whether anywhere in the set of explanatory variables there is some help in accounting for variation in response. A different question is whether *after accounting for the explanatory contributions of several factors, the remaining ones provide any detectable additional ability to model the response, y.* Equivalently, the question can be phrased as to whether the second set of predictors may be dropped from the full MLR model without statistically detectable degradation in one's ability to account for changes in mean response.

A very effective way of thinking about this problem is in terms of the "full" model with k predictor variables and a "reduced model" with *some number, say p, fewer predictors.* (The reduced model then has $k-p$ explanatory variables.) A test of

$$\mathrm{H}_0\text{:all } p \text{ values } \beta_l \text{ corresponding to variables } x_l \text{ not in the reduced model are } 0 \tag{22}$$

can be based on SSR from two regressions. That is, if SSR_{full} is produced by MLR on all k predictors and SSR_{reduced} is produced by MLR on the subset of $k-p$ predictors in the reduced model, then a "partial F test" of hypothesis (22) can be based on the statistic

$$\begin{aligned} F &= \frac{\left(SSR_{\text{full}} - SSR_{\text{reduced}}\right)/p}{SSE_{\text{full}}/\left(n-k-1\right)} \\ &= \frac{\left(SSE_{\text{reduced}} - SSE_{\text{full}}\right)/p}{SSE_{\text{full}}/\left(n-k-1\right)} \end{aligned}$$

and an $F_{p,(n-k-1)}$ reference distribution. p-values are right tail areas beyond the observed value of F. This can be thought of in terms of judging the statistical significance of the increase in R^2 provided by moving from the reduced to the full model and the test statistic has the representation in terms of R^2 values as

$$F = \frac{\left(R^2_{\text{full}} - R^2_{\text{reduced}}\right)/p}{\left(1-R^2_{\text{full}}\right)/\left(n-k-1\right)}$$

The computation of the partial F statistic can be summarized in (and additional intuition provided by) an ANOVA table. The appropriate expansion of Table 5 is presented as Table 6.

Table 6: ANOVA Table for Mulitple Linear Regression Partial F Test

ANOVA Table for Testing H_0:all values β_l corresponding to variables x_l not in the reduced model are 0

Source	SS	df	MS	F
Regression (full)	SSR_{f}	k		
Regression (reduced)	SSR_{r}	$k-p$		
Regression (full|reduced)	$SSR_{\text{f}} - SSR_{\text{r}}$	p	$\left(SSR_{\text{f}} - SSR_{\text{r}}\right)/p$	$MSR_{\text{f}\mid\text{r}}/MSE_{\text{f}}$
Error	SSE_{f}	$n-k-1$	$SSE_{\text{f}}/\left(n-k-1\right)$	
Total	$SSTot$	$n-1$		

30 Some Issues of Interpretation/Use of MLR Inferences

(Text Reference/Reading: V&J Sections 4.2.2,4.2.3,9.2.5)

MLR is a powerful technology. It is also frequently misunderstood/misused by naive analysts. We here make some comments aimed at warning users away from common misinterpretations.

For one thing, it is tempting to treat a coefficient β_l (or its estimate b_l) as "the" effect of x_l on y and to correspondingly treat a large p-value for H_0:$\beta_l = 0$ as evidence that "x_l is of no use in accounting for changes in mean y." But the situation is far mode subtle than that simple phrase suggests.

It is, for example, quite possible to have large p-values for testing *both*

$$\mathrm{H}_0\text{:}\beta_1 = 0 \text{ and } \mathrm{H}_0\text{:}\beta_2 = 0$$

in a 2-variable MLR model including x_1 *and* x_2, and at the same time have small p-values for testing H_0:$\beta_1 = 0$ in SLR on x_1 and H_0:$\beta_2 = 0$ in SLR on x_2. This not inconsistent. The large p-values are indicative that in the presence of the other variable, the variable in question doesn't add significantly to the ability to model changes in y. The small p-values say that if all one has is one of the 2 predictors, it cannot be discarded as useless in predicting y.

The kind of circumstance under which this can occur is one of "**multi-collinearity**." This is the case where one or more approximate linear relationships exist among the predictors $x_1, x_2, \ldots, x_k$ (a situation where the points $(x_{1i}, x_{2i}, \ldots, x_{ki})$ are nearly confined to some hyperplane in $\Re^k$ of dimension lower than k). Consider, for example, a case where $x_{1i} \approx x_{2i}$ for all i. Here, if y is approximately linearly related to x_1, it is equally approximately related to x_2. But *both* variables are not needed in order to model changes in y. Only one or the other is required. — Where there are important sample correlations between predictors, there is multi-collinearity, and it is impossible to cleanly separate the impacts of the various explanatory variables on the mean response.

It is also important to remember that MLR employs exactly the form (20). It is perfectly possible that instead of a form linear in an x_l, something more complicated is needed to describe its relationship to mean y. y might not be linearly related to x_l, but for example be related to a quadratic or a sinusoidal function of x_l. So strictly speaking, inferences about β_l concern only *linear* effects of x_l.

Where a MLR model *does* fit a dataset (e.g. as measured by a large value of R^2) it is important to emphasize that what has been established is a *predictive* relationship, not necessarily a *causal* relationship. The former is a computational fitting matter. The latter concerns physical/real world considerations. And examples abound making it plainly silly to naively assume that "correlation is causation."

Finally, it is important to say that strictly speaking, one really only learns about how y is related to predictors *at those points* $(x_{1i}, x_{2i}, \ldots, x_{ki}) \in \Re^k$ *where one has data.* Anything else is really an *extrapolation.* While using a fitted equation to make extrapolations is common and often practically useful, it is justified only on the basis of *subject matter considerations* outside the kind of purely mathematical and computational ones described in this outline. One must have substantive reasons to believe that the kind of relationship between explanatory variables and mean response seen in the dataset analyzed will extend to points at which one hopes to extrapolate.

31 Some Qualitative Issues/Considerations in Building Models and Predictors for y (Using MLR or Other Methods)

(Text Reference/Reading: V&J Sections 4.1.3&4.2)

Multiple linear regression opens formal discussion of the possibility of seeking functional forms involving *multiple* explanatory factors to describe a response. The last part of 587 will be concerned with methods beyond MLR that are especially helpful in predicting y from inputs $x_1, x_2, \ldots, x_k$ in "big data" situations where one or both of n (the number of data cases) and k (the number of predictors/quantitative explanatory factors) are large. Whether "big data" or "small data" are involved, whether the method is MLR or something else, there are some qualitative points to be made that hold true across all efforts to find approximate relationship between k inputs and an output or response, y. Some of these are the subject of this section.

Good multifactor statistical modeling provides fitted values $\hat{y}_i$ (depending upon $x_{1i}, x_{2i}, \ldots,$ x_{ki}) that effectively approximate observed values y_i. Where full probability models are posited, their assumptions need to be plausible (particularly where one is going to depend substantially upon them for the making of prediction and tolerance intervals). In many (but not all) engineering and physical science contexts, parameters of fitted models have important subject matter interpretations and in those situations, it is important to search for models that are simple and facilitate understanding of the roles of inputs in determining the response. And one hopes to not fail to recognize the effects of important/helpful explanatory variables.

The plotting of residuals is a main tool in achieving these objectives. Normal plots of residuals are helpful in examining "normal errors" model assumptions. Plots of residuals against explanatory variables included in the modeling *and against variables not employed in the modeling* serve as tools for identifying missed opportunities for improving model effectiveness. (In the simplest possible case, where y depends in approximately quadratic fashion upon x, residuals from SLR plotted against x will show a curved pattern. Or, where some variable not taken into account has a strong linear effect on y, residuals plotted against it will show a linear trend.) Plots of residuals against $\hat{y}$ (and ones against the various predictor variables) can be helpful in spotting clear violations of "constant error variance" model assumptions and "outlier" data vectors that simply do not fit with the majority of the n in hand and need careful examination for potential special causes (including things as simple as data-recording blunders).

It is clearly possible to start with a predictor x (or several predictors, $x_1, \ldots, x_k$) and make from it (from them) a new predictor by pugging them into some function. As a simple example, one might start with a single predictor, x, and make from it several more predictors x^2, x^3, and x^4. Then applying MLR ideas to the predictors

$$x_1 = x, x_2 = x^2, x_3 = x^3, \text{ and } x_4 = x^4$$

one can fit the polynomial relationship

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$$

Or beginning from x_1 and x_2 one can make a new predictor $x_3 = x_1 x_2$ and use it in modeling And so on.

In classical treatments of regression analysis, this idea of making new predictor variables from existing ones is usually called making "**transformations**" of predictors. In modern predictive analytics/data mining contexts, it is often termed "**feature engineering**." Whatever it is called, it greatly extends the usefulness of modeling methodologies by adding flexibility to the classes of functions that can be fit to a set of n data vectors $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$.

Even beyond the notion of making fixed transformations of explanatory variables is the possibility of using so called "**smoothing**" methods in large n contexts to more or less automatically make from an explanatory variable (or several explanatory variables) a particularly effective new explanatory variable. (This "automatic choice of a good transformation" is a topic for the end of the course.)

The variety of ways that explicitly (because k one starts with is big) or implicitly (because it is always possible to engineer "new" predictors from "old" ones) there are almost always many potential predictors to employ in modeling a response makes it essential to consider the issue of "**overfitting**" when assessing the quality of a fitted approximate relationship between y and available explanatory variables. This is the possibility that a fitted relationship does a good job of describing data in hand, *but extrapolates very poorly.* For example, one can essentially always improve R^2 (reduce SSE) by adding additional predictors to a MLR ... but that improvement may actually substantially degrade the fitted relationship's usefulness for predicting a new value of y. The hard truth is this:

> The likely effectiveness of a fitted relationship between explanatory variables and a response for predicting response for a new case cannot be assessed by using it to predict responses in the dataset used to make the fit. Rather, one must find a way to test a fitted form on prediction for data *not used in fitting that form.*

A large value of R^2/small residuals is not proof positive that a fitted form is really "any good." The next section in this outline discusses an important methodology addressing this problem, so-called "**cross-validation**."

32 Assessing Prediction Performance by Cross-Validation

(Text Reference/Reading: JWH&T Section 5.1)

We have said that to reliably assess the likely effectiveness of a fitted form in prediction, *one must evaluate prediction performance on cases not used in fitting that model or predictor.* The best available technology for handling this truth is so-called **K-fold cross-validation**. The idea is that one first randomly divides a dataset of size n into K sets of (as nearly as is possible) n/K cases (that we will here call "folds"). Then for each fold, j, one fits the form of interest to all data *except cases in that fold.* For purposes of fixing this idea, for an input vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$ let $\hat{f}^j(\boldsymbol{x})$ be the predictor obtained by fitting to all cases *except those in fold j.* Then for all cases i in fold j, the predicted value of response at $\boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})$,

$$\hat{f}^j(\boldsymbol{x}_i) \ ,$$

is a prediction made based on fitting without case i. A version of an error sum of squares based on cross-validation is then

$$CVSSE = \sum_{j=1}^{K} \sum_{i \text{ in fold } j} \left(y_i - \hat{f}^j(\boldsymbol{x}_i) \right)^2$$

and one might term

$$\frac{1}{n} CVSSE = CVMSPE$$

a (cross-validation mean squared) "prediction error." A measure that is on the same scale as y is the root mean squared prediction error

$$CVRMSPE = \sqrt{CVMSPE}$$

One can hope to reliably assess prediction effectiveness using this metric and thus be in a position to reliably compare different possible fitted forms. Notice that while cross-validation is appropriate for choosing between forms of predictors, once a choice has been made, fitting to the entire dataset in hand is appropriate for purposes of post-fitting prediction.

A natural question is "What should K be?" Typically K in the 5 to 10 range is used. The case $K = n$ is the case where one-at-a-time, all data cases are withheld from fitting and their responses predicted using all other cases. This is often called "leave one out" or LOO cross-validation. While this possibility might seem most natural, there are good technical reasons why $K = 10$ is more common and generally expected to be more reliable.

A second issue that arises is that since the value of $CVRMSPE$ depends upon the random result of splitting of cases into folds, when computationally feasible, it is common to repeat the cross-validation multiple times and replace a single version of the prediction error with an average of multiple values from different random splits into folds. The `caret` package in `R` is an effective tool in implementing cross-validation and, in particular, repeated (and averaged) cross-validation.

33 Logistic Regression (0/1 Responses)

(Text Reference/Reading: V&J Sections A.5.1&A.5.3 Example 19, Section 4.3 JWH&T)

This material concerns modeling and inference for a binomial success probability, p, that is a function of one or more predictors $x_1, x_2, \ldots, x_k$. It is a kind of "regression" like MLR, but is handled with a different methodology. The most common version of this modeling is that where the log odds are taken to be linear in the predictors, i.e. where

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

This is equivalent to a model assumption that

$$p = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\right)} \tag{23}$$

To aid understand the meaning of assumption (23), Figure 1 provides a plot of the "s-shaped" function

$$p\left(u\right) = \frac{\exp\left(u\right)}{1 + \exp\left(u\right)}$$

The assumption (23) says that the input values and parameters combine in linear fashion (as in MLR) to produce the value $u = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ that gets translated into a probability through $p\left(u\right)$. In the case of $k = 1$, $p\left(x\right)$ increases to the right when $\beta_1 > 0$, decreases to the right when $\beta_1 < 0$, has a "steep" plot when $|\beta_1|$ is large, and is .5 exactly when $\beta_0 + \beta_1 x = 0$, i.e. at $x = -\beta_0/\beta_1$.

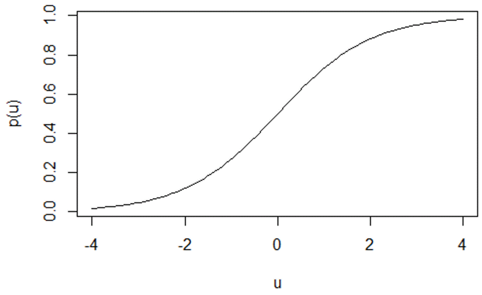

Figure 1: Basic Logistic Curve

We consider a model where for $i = 1, 2, \ldots, n$ independent binomial random variables y_i have corresponding success probabilities

$$p_i = p\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\right)$$

While all that follows can be easily generalized to cases where the numbers of trials for the y_i are larger than 1, for ease of exposition, we'll suppose here that all y_i are based on a single trial each. In this case, the joint pmf of the observables $y_1, y_2, \ldots, y_n$ is the function of the parameters $\beta_0, \beta_1, \ldots, \beta_k$

$$f\left(y_1, y_2, \ldots, y_m | \beta_0, \beta_1, \ldots, \beta_k\right) = \prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1-y_i}$$

With observed values of the y_i plugged into f, one has a function of the parameters only. The logarithm of this is the so-called "log-likelihood function"

$$\begin{aligned} L\left(\boldsymbol{\beta}\right) &= \ln\left(\prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1-y_i}\right) \\ &= \sum_{\substack{i \text{ s.t.} \\ y_i = 1}} \ln p\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\right) + \sum_{\substack{i \text{ s.t.} \\ y_i = 1}} \ln\left(1 - p\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}\right)\right) \end{aligned}$$

that is the basis of inference for the parameter vector $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_k\right)$ and related quantities.

The parameter vector $\boldsymbol{b} = \left(b_0, b_1, \ldots, b_k\right)$ that optimizes (maximizes) $L\left(\boldsymbol{\beta}\right)$ is called the "maximum likelihood estimate" of $\boldsymbol{\beta}$. Further, the shape of the log-likelihood function near the maximum likelihood estimate ($\boldsymbol{b}$) provides confidence regions for the parameter vector $\boldsymbol{\beta}$ and intervals for its entries β_l.

First, the set of parameter vectors $\boldsymbol{\beta}$ with "large" log-likelihood form a confidence set for $\boldsymbol{\beta}$. In fact, for U an upper percentage point of the χ^2_{m-k-1} distribution, those with

$$L\left(\boldsymbol{\beta}\right) > L\left(\boldsymbol{b}\right) - \frac{1}{2}U$$

(those $\boldsymbol{\beta}$ with log-likelihood within $U/2$ of the maximum possible value) form an approximate confidence region (in $\Re^{k+1}$) for $\boldsymbol{\beta}$.

Second, the curvature of the log-likelihood function at the maximizer $\boldsymbol{b}$ provides standard errors for the entries of $\boldsymbol{b}$. That is, for

$$\underset{(k+1)\times(k+1)}{\boldsymbol{H}} = \left[\frac{\partial^2}{\partial\beta_i \partial\beta_j} L\left(\boldsymbol{\beta}\right)\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{b}}$$

the "Hessian" matrix (the matrix of second partials of the log-likelihood at the maximizer $\boldsymbol{b}$), estimated variances of the entries of $\boldsymbol{b}$ can be obtained as diagonal entries of

$$-\boldsymbol{H}^{-1}$$

(the negative inverse Hessian). The square roots of these then serve as standard errors for the estimated coefficients b_l (values se_{b_l}) that get printed out by statistical systems like `R`. Corresponding approximate confidence limits for β_l are then

$$b_l \pm z\text{se}_{b_l}$$

Somewhat more reliable confidence limits can be produced by a more complicated/subtle method and can be gotten from `glm()`. (One finds all values β_l^* for which there is a $\boldsymbol{\beta} \in \Re^{k+1}$ with $\beta_l = \beta_l^*$ and $L\left(\boldsymbol{\beta}\right) > L\left(\boldsymbol{b}\right) - \frac{1}{2}U$ for U an upper percentage point of χ^2_1.)

The value

$$\hat{u} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

(parallel to $\hat{y}$ in display (18)) serves as a **fitted log odds**. The `glm()` package in `R` will produce fitted values for the dataset (and for new vectors of inputs) and will also produce corresponding standard errors. Call these $\text{se}_{\hat{u}}$. Then, approximate confidence limits for the log odds $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$ are

$$\hat{u} \pm z \cdot \text{se}_{\hat{u}}$$

Simply inserting these limits into the function $p\,(u)$ produces confidence limits for the success probability at inputs $x_1, x_2, \ldots, x_k$, namely

$$p\,(\hat{u} - z \cdot \text{se}_{\hat{u}}) \quad \text{and} \quad p\,(\hat{u} + z \cdot \text{se}_{\hat{u}})$$

giving a way to see how much one knows about the success probabilities at various vectors of inputs.

34 Non-Linear Regression

(Text Reference/Reading:)

We consider the generalization of multiple linear regression involving p predictor variables $x_1, x_2, \ldots, x_p$ and k (unknown) parameters $\beta_1, \beta_2, \ldots, \beta_k$ (that, as convenient, we will assemble into vectors $\boldsymbol{x}$ and $\boldsymbol{\beta}$ respectively). We assume that there is some known function $f(\boldsymbol{x}; \boldsymbol{\beta})$ that provides the mean value of an observable variable, y, in terms of these. Then, as in MLR, we assume that for independent mean 0 and variance σ^2 normal variables ϵ_i, for $i = 1, 2, \ldots, n$

$$y_i = f(\boldsymbol{x}_i; \boldsymbol{\beta}) + \epsilon_i \tag{24}$$

Notice that in model (24), exactly as in MLR, there is an assumption that for inputs $x_1, x_2, \ldots, x_p$ the distribution of y around the mean $\mu_{y|x_1,x_2,\ldots,x_p} = f(\boldsymbol{x}; \boldsymbol{\beta})$ is normal with a standard deviation σ that doesn't depend upon the inputs. The innovation here (relative to MLR) is simply the possibility that $f(\boldsymbol{x}; \boldsymbol{\beta})$ doesn't have the MLR form $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$. Particularly in the physical sciences and engineering, theories (e.g. involving differential equations) and well-established empirical relationships provide other favorite functional forms. Some of these forms don't even have explicit representations and are simply defined as "the solution to a system of differential equations." This section is about how they can to some extent be handled statistically in a way parallel to the handling of MLR.

First, there is the question of how to process n data vectors into estimates of the parameters σ and $\boldsymbol{\beta}$. Just as in MLR, one can use least squares, i.e. minimize

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i; \boldsymbol{\beta}))^2$$

Conceptually, this is exactly as in MLR (except for the fact that the function $S(\boldsymbol{\beta})$ is not a quadratic function of $\boldsymbol{\beta}$). Operationally (because S is not so simple) one must employ iterative algorithms to search for an optimizer. In `R` one can use the `nls()` routine instead of the `lm()` routine. Further, because statistical theory for the general model (24) is not as clean as for the special case of MLR, only approximate methods of inference can be identified, and they are impossible to describe completely at a Stat 587 level of background. What will be done in the balance of this section is to try to give some understandable description of and motivation for what is possible and is implemented in good statistical packages.

So, without getting into the numerical details of exactly what algorithms are used to locate it, suppose that $\boldsymbol{b}$ is an optimizer of $S(\boldsymbol{\beta})$, that is

$$S(\boldsymbol{b}) = \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta})$$

($\boldsymbol{b}$ minimizes an "error sum of squares") and is an "ordinary least squares estimate" of $\boldsymbol{\beta}$. Of course, entries of $\boldsymbol{b}$ serve as estimates of the entries of $\boldsymbol{\beta}$.

The (minimum) sum of squares corresponding to $\boldsymbol{b}$, namely

$$SSE = S(\boldsymbol{b}) = \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i; \boldsymbol{b}))^2$$

can (as in MLR) form a basis for estimating σ. In particular, a simple estimate of σ^2 is

$$\widehat{\sigma^2} = \frac{SSE}{n-k}$$

Simply carrying over ideas from MLR, very approximate confidence limits for σ are

$$\hat{\sigma}\sqrt{\frac{n-k}{U}} \text{ and } \hat{\sigma}\sqrt{\frac{n-k}{L}}$$

for U and L small upper and lower percentage points of the χ^2_{n-k} distribution. More subtle methods than this are available and are sometimes implemented in non-linear least squares software.

Confidence *regions* for locating the whole parameter vector $\boldsymbol{\beta}$ are sometimes of interest. Carrying over an idea from MLR (that wasn't discussed in the context of MLR but does work there exactly) one can use as a confidence region a set of parameters $\boldsymbol{\beta}$ for which $S(\boldsymbol{\beta})$ is not much larger than the minimum value, $S(\boldsymbol{b})$. In particular, the set of parameters $\boldsymbol{\beta}$ for which

$$S(\boldsymbol{\beta}) \leq S(\boldsymbol{b})\left(1 + \frac{k}{n-k}U\right)$$

for U a small upper percentage point of the $F_{k,n-k}$ distribution serves as an approximate confidence region for $\boldsymbol{\beta}$.

Standard errors for the b_l can be obtained in much the same way as they are in logistic regression, based on the curvature of an appropriate log-likelihood function (involving the inversion of a Hessian, etc.). These values se_{b_l} are printed out on most non-linear regression outputs and approximate confidence limits for β_l are

$$b_l \pm t\text{se}_{b_l}$$

for t a small upper percentage point of the t_{n-k} distribution.

More reliable approximate confidence limits for individual coefficients can be made via a method related to the method for making confidence regions for $\boldsymbol{\beta}$ discussed above. That is, the set of parameters β_l^* for which there is a $\boldsymbol{\beta} \in \Re^k$ with $\beta_l = \beta_l^*$ and

$$S(\boldsymbol{\beta}) \leq S(\boldsymbol{b})\left(1 + \frac{1}{n-k}U\right)$$

for U a small upper percentage point of the $F_{1,n-k}$ distribution serves as an approximate confidence region for β_l. This method is sometimes implemented in non-linear regression programs as an alternative to the t intervals.

It is also possible to make confidence limits for the mean response $f(\boldsymbol{x};\boldsymbol{\beta})$, but no version of this seems to be presently implemented in `nls()`. In particular, no standard errors for fits (the $\text{se}_{\hat{y}}$) seem to be implemented at the moment. If they were, then approximate prediction limits for a next y at a particular set of conditions would be

$$f(\boldsymbol{x};\boldsymbol{b}) \pm t\sqrt{\widehat{\sigma^2} + (\text{se}_{\hat{y}})^2}$$

35 (Complete) Two-Way Factorial Analyses

(Text Reference/Reading: V&J Sections 4.3.1-4.3.2 and 8.1.1-8.1.3) (See also V&J SMQA Section 5.2)

We now begin to consider modeling and statistical analysis for circumstances where a mean response potentially depends upon several factors which may not be quantitative in nature (and thus MLR and its extensions are not obviously applicable). We begin with the simplest case, where two factors, that we will here simply call "A" and "B," have respectively I and J possible "levels" (these are different settings or values of the two factors), and a dataset has at least one observation for each of the $I \cdot J$ different combinations of a level of A with a level of B. This kind of circumstance is called a *complete* (because no combinations lack data) *two-way factorial* context.

Typical data analysis here is supported by the usual one-way normal model (9) of Section 20 rewritten in a way that makes explicit the natural two-way structure in the $r = IJ$ different conditions under study. That is, with

$$y_{ijk} = \text{the } k\text{th observation at level } i \text{ of Factor A and level } j \text{ of Factor B}$$

for $i = 1, \ldots, I, j = 1 \ldots, J$, sample sizes n_{ij}, and $k = 1, \ldots, n_{ij}$ for each i, j pair, the model is

$$y_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

for iid mean 0 variance σ^2 random errors ϵ_{ijk} and $r = IJ$ means μ_{ij}. (The model parameters are the means and the single standard deviation.)

We will employ "dot subscript" notation for averages of various sample means and parameters. That is, we let

$$\bar{y}_{ij} = \frac{1}{n_{ij}} \sum_{k=1}^{n_{ij}} y_{ijk}, \quad \bar{y}_{i.} = \frac{1}{J} \sum_{j=1}^{J} \bar{y}_{ij}, \quad \bar{y}_{.j} = \frac{1}{I} \sum_{i=1}^{I} \bar{y}_{ij}, \quad \text{and} \quad \bar{y}_{..} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \bar{y}_{ij}$$

and

$$\mu_{i.} = \frac{1}{J} \sum_{j=1}^{J} \mu_{ij}, \quad \mu_{j.} = \frac{1}{I} \sum_{i=1}^{I} \mu_{ij}, \quad \text{and} \quad \mu_{..} = \frac{1}{IJ} \sum_{i=1}^{I} \sum_{j=1}^{J} \mu_{ij}$$

Sound understanding of patterns or structure in the mean responses μ_{ij} can begin by making so-called "interaction plots" of the sample means, $\bar{y}_{ij}$. These are made by plotting $\bar{y}_{ij}$ against i (level of Factor A) or against j (level of Factor B), connecting consecutive plotted means for a given j in the first case or i in the second with line segments. These plots can be enhanced using "error bars" around $\bar{y}_{ij}$ derived from confidence limits for μ_{ij},

$$\bar{y}_{ij} \pm t \frac{s_{\text{P}}}{\sqrt{n_{ij}}}$$

It is also possible and important to define so-called (theoretical and fitted/estimated) "main effects" for the factors individually and (theoretical and fitted/estimated) "two-factor interactions" for the factors in a two-way factorial. The main effects for Factor A at level i are

$$\alpha_i = \mu_{i.} - \mu_{..} \text{ (model/theoretical)} \quad \text{and} \quad a_i = \bar{y}_{i.} - \bar{y}_{..} \text{ (fitted/estimated/empirical)}$$

The main effects for Factor B at level j are

$$\beta_j = \mu_{.j} - \mu_{..} \ \ \text{(model/theoretical)} \ \ \text{and} \ \ b_j = \bar{y}_{.j} - \bar{y}_{..} \ \ \text{(fitted/estimated/empirical)}$$

And the two-factor interactions for Factors A and B at combination i, j are

$$\alpha\beta_{ij} = \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j) \ \ \text{(model/theoretical)}$$
$$\text{and} \ \ ab_{ij} = \bar{y}_{ij} - (\bar{y}_{..} + a_i + b_j) \ \ \text{(fitted/estimated/empirical)}$$

Main effects measure the difference between level average means and an overall average mean. Two-factor interactions measure differences between combination/cell means and what can be accounted for in terms of an overall mean and main effects. These latter are a kind of measure of "dependence of a factor's effect upon the level of the other factor." When they are negligible, "interaction plots" have a kind of parallelism property, whereby means plotted against level of (say) A move up and down similarly for each level of (say) B.

A basic kind of inference available for main effects and two-factor interactions derives from the fact that theoretical/model effects (main and interaction alike) are "L's" (see display (10)) and fitted/estimated effects are corresponding "$\hat{L}$'s" (see display (11)) of Section 21. This means that formula (12) gives confidence limits for α_i's, β_j's, and $\alpha\beta_{ij}$'s. In addition, the same is then true for differences in main effects, $\alpha_i - \alpha_{i'} = \mu_{i.} - \mu_{i'.}$ or $\beta_j - \beta_{j'} = \mu_{.j} - \mu_{.j'}$ that measure differences in level average mean responses. That is, for

$$L = \text{a model effect or difference in main effects}$$

one has

$$\hat{L} = \text{the corresponding fitted effect or difference in fitted main effects}$$

and confidence limits

$$\hat{L} \pm t s_{\text{P}} \sqrt{\sum_{i,j} \frac{c_{ij}^2}{n_{ij}}}$$

The only potential mystery is the value of the sum under the square root above. While this can be worked out from first principles if one identifies the coefficients applied to each combination mean in order to make L, Tables 8.3 and 8.4 on pages 556 and 557 of V&J give formulas for this sum in both "balanced data" cases (where all n_{ij} are some common value m) and in general.

36 MLR and Two-Way Factorial Analyses (and ANOVA)

(Text Reference/Reading: V&J Section 9.3.2)

In a way that is perhaps initially quite surprising, MLR enables more detailed analyses of 2-way factorial data than what is provided in the previous section. That is built on a clever "coding" idea that represents I levels of Factor A with $I-1$ "dummy" variables and J levels of Factor B with $J-1$ other dummy variables and the facts that as defined in Section 35,

1. A main effects sum to 0, $\sum_{i=1}^{I}\alpha_i = 0$,
2. B main effects sum to 0, $\sum_{j=1}^{J}\beta_j = 0$, and
3. AB two-factor interactions sum to 0 across levels of either factor, $\sum_{j=1}^{J}\alpha\beta_{ij} = 0$ for each i and $\sum_{i=1}^{I}\alpha\beta_{ij} = 0$ for each j.

The basic idea is this. For $i = 1, 2, \ldots, I-1$ define

$$x_i^{\text{A}} = \begin{cases} 1 & \text{if the case is from level } i \text{ of Factor A} \\ -1 & \text{if the case is from level } I \text{ of Factor A} \\ 0 & \text{otherwise} \end{cases} \qquad (25)$$

Similarly, for $j = 1, 2, \ldots, J-1$ define

$$x_j^{\text{B}} = \begin{cases} 1 & \text{if the case is from level } j \text{ of Factor B} \\ -1 & \text{if the case is from level } J \text{ of Factor B} \\ 0 & \text{otherwise} \end{cases} \qquad (26)$$

Then as it turns out, in a MLR mean expression including all $I-1$ "predictors" x_i^{A}, all $J-1$ "predictors" x_j^{B} , and all $(I-1)(J-1)$ "product predictors" $x_i^{\text{A}}x_j^{\text{B}}$ (including therefore $IJ-1$ "predictors" overall) each μ_{ij} has a different representation in terms of the regression coefficients (the β_l's). Further,

1. the "intercept" "β_0" is in fact $\mu_{..}$,
2. each regression coefficient "β_l" corresponding to an x_i^{A} is the corresponding main effect α_i as defined in Section 35,
3. each regression coefficient "β_l" corresponding to an x_j^{B} is the corresponding main effect β_j as defined in Section 35, and
4. each regression coefficient "β_l" corresponding to a product $x_i^{\text{A}}x_j^{\text{B}}$ is the corresponding two-factor interaction $\alpha\beta_{ij}$ as defined in Section 35.

(This is illustrated in explicit terms for the 3 × 3 case in Section 9.3.2 of V&J.)

So, one can do two-factor factorial inference (including estimation of the main effects and interactions) using (either explicitly or behind the scenes) MLR computations. This includes the kind of estimation of effects and differences in main effects considered in the previous section. But beyond this, there is the possibility of considerations based on "reduced" models (where interactions and perhaps one kind of main effects are all 0). Some of the things that can be done are the subject of the rest of this section.

The hypothesis that there are no interactions in a two-way set of means, H_0:all $\alpha\beta_{ij} = 0$, can be phrased in MLR terms as H_0:all "β_l" corresponding to products $x_i^{\mathrm{A}} x_j^{\mathrm{B}}$ are 0, and tested using the "full model/reduced model" paradigm employing the statistic

$$F = \frac{\left(SSR\left(\text{all } x_i^{\mathrm{A}}\text{,all } x_j^{\mathrm{B}}\text{, all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) - SSR\left(\text{all } x_i^{\mathrm{A}}\text{,all } x_j^{\mathrm{B}}\right)\right) / \left(I-1\right)\left(J-1\right)}{SSE\left(\text{all } x_i^{\mathrm{A}}\text{,all } x_j^{\mathrm{B}}\text{, all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) / \left(n - IJ\right)}$$

with an $F_{(I-1)(J-1),(n-IJ)}$ reference distribution. (Notice that the null hypothesis here is that $\mu_{ij} = \mu_{..} + \alpha_i + \beta_j$ for all i, j, i.e. that a "no interactions model" describes the influence of the two factors on the mean response.)

Similarly, the hypothesis that there are no interactions and no B main effects, H_0:all $\alpha\beta_{ij} = 0$ and all $\beta_j = 0$, can be phrased in MLR terms as H_0:all "β_l" corresponding to x_j^{B} and to products $x_i^{\mathrm{A}} x_j^{\mathrm{B}}$ are 0, and tested using the "full model/reduced model" paradigm employing the statistic

$$F = \frac{\left(SSR\left(\text{all } x_i^{\mathrm{A}}\text{,all } x_j^{\mathrm{B}}\text{, all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) - SSR\left(\text{all } x_i^{\mathrm{A}}\right)\right) / I\left(J-1\right)}{SSE\left(\text{all } x_i^{\mathrm{A}}\text{,all } x_j^{\mathrm{B}}\text{, all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) / \left(n - IJ\right)}$$

with an $F_{I(J-1),(n-IJ)}$ reference distribution. (The null hypothesis here is that $\mu_{ij} = \mu_{..} + \alpha_i$ for all i and j, i.e. that an "A main effects only model" describes the influence of the two factors on the mean response.)

These tests, ANOVA tables that summarize them, and the sums of squares that they are built on are typically output automatically by "ANOVA" or factorial analysis routines (without any requirement that a user actually set up the dummy variables that stand behind them). But there are some subtleties that must be understood when using these.

When two-way full factorial data are balanced (all n_{ij} are the same)

$$SSR\left(\text{all } x_i^{\mathrm{A}}\right) = SSR\left(\text{all } x_i^{\mathrm{A}} | \text{all } x_j^{\mathrm{B}}\right) = SSR\left(\text{all } x_i^{\mathrm{A}} | \text{all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) = SSR\left(\text{all } x_i^{\mathrm{A}} | \text{all } x_j^{\mathrm{B}}\text{, all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right)$$

and

$$SSR\left(\text{all } x_j^{\mathrm{B}}\right) = SSR\left(\text{all } x_j^{\mathrm{B}} | \text{all } x_i^{\mathrm{A}}\right) = SSR\left(\text{all } x_j^{\mathrm{B}} | \text{all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) = SSR\left(\text{all } x_j^{\mathrm{B}} | \text{all } x_i^{\mathrm{A}}\text{, all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right)$$

and

$$SSR\left(\text{all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) = SSR\left(\text{all } x_i^{\mathrm{A}} x_j^{\mathrm{B}} | \text{all } x_i^{\mathrm{A}}\right) = SSR\left(\text{all } x_i^{\mathrm{A}} x_j^{\mathrm{B}} | \text{all } x_j^{\mathrm{B}}\right) = SSR\left(\text{all } x_i^{\mathrm{A}} x_j^{\mathrm{B}} | \text{all } x_i^{\mathrm{A}}\text{, all } x_j^{\mathrm{B}}\right)$$

In this context, it then makes sense to write

$$\begin{aligned} SSA &= SSR\left(\text{all } x_i^{\mathrm{A}}\right) \\ SSB &= SSR\left(\text{all } x_j^{\mathrm{B}}\right) \\ SSAB &= SSR\left(\text{all } x_i^{\mathrm{A}} x_j^{\mathrm{B}}\right) \end{aligned}$$

and have

$$SSA + SSB + SSAB = SSTr$$

That is, there are natural meanings for "A," "B," and "AB" components of the "treatment" sum of squares from a one-way analysis of the $r = IJ$ samples, providing entries in an ANOVA table like Table 7. The F statistics in this table provide tests for hypotheses that there are no differences among the IJ means, all A main effects are all 0, all B main effects are 0, and all AB interactions are 0 in the "full model" (that allows all combinations to have completely unconstrained values for mean responses).

Table 7: Form of the ANOVA Table for a Two-Way Complete Factorial Analysis

Source	SS	df	MS	F
Treatments	$SSTr$	$IJ-1$	$SSTr/(IJ-1)$	$MSTr/MSE$
A	SSA	$I-1$	$SSA/(I-1)$	MSA/MSE
B	SSB	$J-1$	$SSB/(J-1)$	MSB/MSE
A×B	$SSAB$	$(I-1)(J-1)$	$SSAB/(I-1)(J-1)$	$MSAB/MSE$
Error	SSE	$n-IJ$	$SSE/(n-IJ)$	
Total	$SSTot$	$n-1$		

Where two-way factorial data are not balanced, there is no one obvious partition of $SSTr$ into parts uniquely attributable separately to the two factors and their interactions. One possible partition is

$$\begin{aligned} SSTr = {} & SSR\left(\text{all } x_i^{\text{A}}\right) + \left(SSR\left(\text{all } x_i^{\text{A}}, \text{all } x_j^{\text{B}}\right) - SSR\left(\text{all } x_i^{\text{A}}\right)\right) \\ & + \left(SSR\left(\text{all } x_i^{\text{A}}, \text{all } x_j^{\text{B}}, \text{ all } x_i^{\text{A}} x_j^{\text{B}}\right) - SSR\left(\text{all } x_i^{\text{A}}, \text{all } x_j^{\text{B}}\right)\right) \end{aligned}$$

and when factors are entered in the call of an ANOVA routine in the order "A first and then B" typically the routine will report and ANOVA table using this partition

$$\begin{aligned} \text{"}SSA\text{"} &= SSR\left(\text{all } x_i^{\text{A}}\right) \\ \text{"}SSB\text{"} &= SSR\left(\text{all } x_i^{\text{A}}, \text{all } x_j^{\text{B}}\right) - SSR\left(\text{all } x_i^{\text{A}}\right) \\ \text{"}SSAB\text{"} &= SSR\left(\text{all } x_i^{\text{A}}, \text{all } x_j^{\text{B}}, \text{ all } x_i^{\text{A}} x_j^{\text{B}}\right) - SSR\left(\text{all } x_i^{\text{A}}, \text{all } x_j^{\text{B}}\right) \end{aligned}$$

Notice that only the last of these is an appropriate numerator sum of squares for testing an hypothesis in the full model and that what are reported as "SSA" and "SSB" will typically be different if the factors are entered in the call of the ANOVA routine in the order "B first and then A." So, in fact, treating a two-way factorial in MLR terms provides the most transparency and control over exactly what is being portrayed in a data analysis.

Subtleties/complications of interpretation introduced into two-way factorial analysis by lack of balance in a dataset extend beyond sums of squares and F tests, to values for fitted effects and predicted values for reduced models (ones that don't include all $IJ-1$ predictors and thus impose constraints on the μ_{ij}). That is, considering the three sets of predictors

$$\text{"all } x_i^{\text{A}}\text{," "all } x_j^{\text{B}}\text{," and "all } x_i^{\text{A}} x_j^{\text{B}}\text{"}$$

the full model includes all three sets, but reduced models can be built using only one or two. **For balanced data cases** MLR fits of the reduced models all produce

1. estimated coefficients agreeing with corresponding ones produced in a fit of the full model (and thus equal to the fitted effects defined in Section 35), and then

2. predicted responses (fitted means) that are sums of the relevant fitted effects defined in Section 35.

But where two-way factorial data are not balanced, these simplifications do not hold. Estimates of "β_l"'s depend upon which model is being fit (what "other" types of predictors are being considered) and while "$\hat{y}$" values are what they always are in MLR, for reduced models they **are typically not** simply sums of the fitted effects defined in Section 35.

37 Complete p Factor Factorial Studies (Generalities)

(Text Reference/Reading: V&J Sections 4.3.3-4.3.4, 8.2.1-8.2.2&9.3.2) (See also V&J SMQA Section 5.3.1)

p-way complete factorial studies consist of at least 1 observation at every combination of levels of p different factors. The ideas of the previous two sections generalize to cover analysis of these data structures, after one makes sensible definitions of factorial effects in these p-factor contexts. In this section we treat the factorial analysis problem, using the $p = 3$ case as the focus of discussion and simply alluding to how the ideas must generalize to $p > 3$.

So suppose that 3 factors that we will here call "A," "B," and "C," have respectively I, J, and K possible levels and that a dataset has at least one observation for each of the $I \cdot J \cdot K$ different combinations of a level of A with a level of B with a level of C. As for the two-way factorial situation, typical data analysis here is supported by the one-way normal model (9) of Section 20 rewritten in a way that makes explicit the natural factorial structure in the $r = IJK$ different conditions under study. Here, with

$y_{ijkl} =$ the lth observation at level i of Factor A, level j of Factor B, and level k of Factor C

for $i = 1, \ldots, I, j = 1 \ldots, J, k = 1, \ldots, K$, sample sizes n_{ijk}, and $l = 1, \ldots, n_{ijk}$ for each i, j, k triple, the model is

$$y_{ijkl} = \mu_{ijk} + \epsilon_{ijkl} \tag{27}$$

for iid mean 0 variance σ^2 random errors ϵ_{ijkl} and $r = IJK$ means μ_{ijk}. (The model parameters are the $r = IJK$ means and the single standard deviation, and an additional subscript beyond the 3 of the previous two sections is required to describe $p = 3$ datasets.)

We continue to employ "dot subscript" notation for averages of means, so that

$$\mu_{ij.} = \frac{1}{K}\sum_{k=1}^{K}\mu_{ijk}, \quad \mu_{i.k} = \frac{1}{J}\sum_{j=1}^{J}\mu_{ijk}, \quad \mu_{.jk} = \frac{1}{I}\sum_{i=1}^{I}\mu_{ijk}, \quad \mu_{i..} = \frac{1}{JK}\sum_{j,k}\mu_{ijk},$$
$$\mu_{.j.} = \frac{1}{IK}\sum_{i,k}\mu_{ijk}, \quad \mu_{..k} = \frac{1}{IJ}\sum_{i,j}\mu_{ijk}, \quad \text{and} \quad \mu_{...} = \frac{1}{IJK}\sum_{i,j,k}\mu_{ijk}$$

Then, in analogy to the two-way case, main effects in a three-way factorial are differences between average means at a level of a single factor of interest and the overall average mean,

$$\begin{aligned}\alpha_i &= \mu_{i..} - \mu_{...}\\ \beta_j &= \mu_{.j.} - \mu_{...}\\ \gamma_k &= \mu_{..k} - \mu_{...}\end{aligned}$$

Two-way interactions are differences between average means at a combination of levels of two factors of interest and what can be accounted for by the overall mean and the two corresponding main effects,

$$\begin{aligned}\alpha\beta_{ij} &= \mu_{ij.} - (\mu_{...} + \alpha_i + \beta_j)\\ \alpha\gamma_{ik} &= \mu_{i.k} - (\mu_{...} + \alpha_i + \gamma_k)\\ \beta\gamma_{jk} &= \mu_{.jk} - (\mu_{...} + \beta_j + \gamma_k)\end{aligned}$$

Finally, three-way interactions are differences between combination means and what can be accounted for by the overall mean, main effects, and two-factor interactions

$$\alpha\beta\gamma_{ijk} = \mu_{ijk} - (\mu_{...} + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk})$$

(In higher way factorials, interactions of a given order are differences between average means and what can be accounted for by the overall mean, main effects, and interactions of order lower than the ones being defined.) Typical data analysis in a three-way factorial then concerns determining which of these effects are important and making subject matter interpretations.

In the next section we will consider factorial analyses where every factor has only 2 levels. There, some very nice simplifications of formulas, notations, and interpretations are possible. Here we say what can be done in general using the "coding in MLR" ideas of the last section.

Again defining dummy variables for Factors A and B as in displays (25) and (26), now define as well dummy variables for Factor C. For $k = 1, 2, \ldots, K-1$ let

$$x_k^{\mathrm{C}} = \begin{cases} 1 & \text{if the case is from level } k \text{ of Factor C} \\ -1 & \text{if the case is from level } K \text{ of Factor C} \\ 0 & \text{otherwise} \end{cases} \tag{28}$$

Then a MLR version of the full model (27) includes all sets of predictors

$$\text{"all } x_i^{\mathrm{A}}\text{," "all } x_j^{\mathrm{B}}\text{," "all } x_k^{\mathrm{C}}\text{," "all } x_i^{\mathrm{A}}x_j^{\mathrm{B}}\text{," "all } x_i^{\mathrm{A}}x_k^{\mathrm{C}}\text{," "all } x_j^{\mathrm{B}}x_k^{\mathrm{C}}\text{," and "all } x_i^{\mathrm{A}}x_j^{\mathrm{B}}x_k^{\mathrm{C}}\text{"} \tag{29}$$

and a regression coefficient "β_l" corresponding to a dummy variable or product of dummy variables for more than one factor is the corresponding factorial main effect or interaction. (Main effects and interactions involving the "last" level of one or more factors are available as sums and differences of others of the given type across levels of relevant factors, since by definition main effects and interactions sum to 0 across levels of any factor referenced in the effect name.) While such regressions can be set up "by hand" and the form of MLR output thereby carefully controlled by an analyst, it is quite common to instead simply employ a factorial analysis/"ANOVA" routine (with its pre-programmed choices of output format) to do computations. These routines effectively create their own dummy variables and employ MLR computations in the background.

Many different partial F tests based on the full model/reduced model paradigm can be implemented using the sets of predictors (29). In particular, F tests that all effects corresponding to a single set are 0 in the full model (27) can be based on a full model regression involving all predictors (29) and a reduced model where a single set is dropped from the list (29). Exactly what tests a pre-programmed routine will by default enable is something that must be carefully determined by a user.

As in the case of two-way factorial analyses, **balanced data** (where all combinations of levels of the p factors have the same sample sizes) provide great conceptual simplifications. Where all n_{ijk} in a 3-way study are the same, there is an obvious single sum of squares to associate with each set of predictors in display (29). That is, whether or not any other set or sets of predictors from the list is already included in a model for y, adding a particular set will increase the regression sum of squares by the same amount. It thus makes sense to call that a sum of squares associated with the set of effects. For example

$$SSR\left(\text{"all } x_i^{\mathrm{A}}\text{"}\right) = SSR\left(\text{"all } x_i^{\mathrm{A}}\text{," "all } x_j^{\mathrm{B}}\text{," and "all } x_k^{\mathrm{C}}\text{"}\right) - SSR\left(\text{"all } x_j^{\mathrm{B}}\text{" and "all } x_k^{\mathrm{C}}\text{"}\right) = \cdots$$

so that it makes sense to call

$$SSA = SSR\left(\text{"all } x_i^{\mathrm{A}}\text{"}\right)$$

In this relatively simple situation, $SSTr$ based on the one-way model (with $r = IJK$ different conditions) partitions naturally as

$$SSTr = SSA + SSB + SSC + SSAB + SSAC + SSBC + SSABC$$

and this partition is typically shown in an ANOVA table along with degrees of freedom

$$\text{df A} = I-1, \text{df B} = J-1, \text{df C} = K-1, \text{df AB} = (I-1)(J-1), \text{df AC} = (I-1)(K-1),$$
$$\text{df BC} = (J-1)(K-1), \quad \text{and} \quad \text{df ABC} = (I-1)(J-1)(K-1)$$

Further, estimated effects (that are estimated regression coefficients "b_l") for reduced models including only some of the sets of predictors (29) are the same as those for the full model (27). So there is no ambiguity regarding the apparent sizes of the factorial effects related to which other effects are simultaneously considered.

Unbalanced factorial data have conceptually more difficult problems of interpretation. There is no obvious single sum of squares to associate with a given set of predictors in display (29) in $p = 3$ problems. The additional regression sum of squares provided by one of the sets depends upon which other set or sets have already been accounted for (in a reduced model). So while it is common for a factorial analysis routine to output some kind of an ANOVA table, exactly what the sums of squares represent is not necessarily obvious and almost surely depends upon the order factors are listed in when the call to the routine is made. (Typically different sums of squares appear for different orders.) Exactly what full and reduced models are implicit in the form of the ANOVA table must be carefully worked out by an analyst, who needs to be sure that he or she has the correct raw material for whatever comparisons of models is of interest. And the estimated effects provided by such a routine will also typically depend upon the user-specified ordering of the factors. For example, there are thus no obvious "estimated B main effects," rather only "estimated B main in the presence of XXX effects." A user must therefore be very careful and thoughtful in interpreting what a routine provides for output for unbalanced factorial data.

38 Special Methods for 2^p Factorials

(Text Reference/Reading: V&J Sections 4.3.3-4.3.5 and 8.2.1-8.2.3) (See also V&J SMQA Section 5.3)

We now consider the special case of p-way factorial analysis where each of the p factors has only 2 levels — the so-called $2 \times 2 \times \cdots \times 2$ or 2^p studies. There are two reasons for giving special attention to 2-level factors. The first is that there is special notation and structure that make their analysis most transparent. The second is that as a practical matter, one can rarely afford p-factor factorial experimentation with many (more than 2) levels of the factors. As we did in the previous section, we'll phrase the discussion mostly in terms of the $p = 3$ case, counting on the reader to make the natural extensions to $p > 3$.

It is typical in 2^p studies to make an arbitrary choice of one level of each factor as a first or "low" level and the other level as the second or "high" level. Further, it is often useful employ the "$-$" designator for the low level and the "$+$" designator for the high level. In addition it is common to adopt a shorthand naming convention for the 2^p different combinations that calls each by a string of letters corresponding to *those factors appearing in the combination at their 2nd or high levels.* Table 8 summarizes these notational conventions for index i indicating level of A, index j indicating the level of B, and index k indicating the level of C. While the "ijk" notation is perfectly general and could be applied to any $I \times J \times K$ factorial, the $+/-$ notation used in the table and the special "2^p name" convention are special to the present case where every factor has only 2 levels.

Table 8: Naming Convention for Combinations in a 2^p Factorial

A	B	C	2^3 name	i	j	k
$-$	$-$	$-$	(1)	1	1	1
$+$	$-$	$-$	a	2	1	1
$-$	$+$	$-$	b	1	2	1
$+$	$+$	$-$	ab	2	2	1
$-$	$-$	$+$	c	1	1	2
$+$	$-$	$+$	ac	2	1	2
$-$	$+$	$+$	bc	1	2	2
$+$	$+$	$+$	abc	2	2	2

It is often helpful for understanding the results of a 2^3 factorial study to plot the sample means obtained on the corners of a cube as shown in Figure 2.

In a way consistent with the notation of the last few sections we'll let

$$\bar{y}_{ijk} = \frac{1}{n_{ijk}} \sum_{l=1}^{n_{ijk}} y_{ijkl}, \quad \bar{y}_{ij.} = \frac{1}{K} \sum_{k=1}^{K} \bar{y}_{ijk}, \quad \bar{y}_{i.k} = \frac{1}{J} \sum_{j=1}^{J} \bar{y}_{ijk}, \quad \bar{y}_{.jk} = \frac{1}{I} \sum_{i=1}^{I} \bar{y}_{ijk},$$

$$\bar{y}_{i..} = \frac{1}{JK} \sum_{j,k} \bar{y}_{ijk}, \quad \bar{y}_{.j.} = \frac{1}{IK} \sum_{i,k} \bar{y}_{ijk}, \quad \bar{y}_{..k} = \frac{1}{IJ} \sum_{i,j} \bar{y}_{ijk}, \quad \text{and} \quad \bar{y}_{...} = \frac{1}{IJK} \sum_{i,j,k} \bar{y}_{ijk}$$

and note that there are fitted versions of the 2^3 factorial effects defined in the previous section that can be defined in terms of these sample means. (These will agree exactly with what is produced

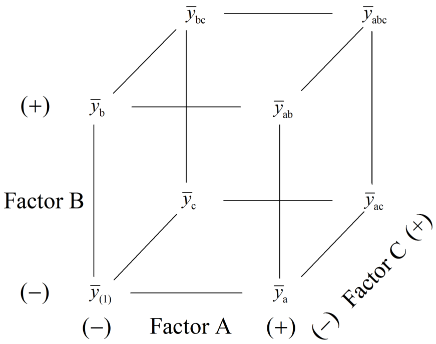

Figure 2: Sample means from a 2^3 factorial

*using **all** sets of dummy variables in display* (29) in an MLR model. The point of providing the formulas here is more for lending intuition about the problem than for recommending their use in practice. In practice, either the MLR ideas or the so-called "Yates algorithm" is typically easier.) That is, fitted main effects are

$$\begin{aligned} a_i &= \bar{y}_{i..} - \bar{y}_{...} \\ b_j &= \bar{y}_{.j.} - \bar{y}_{...} \\ c_k &= \bar{y}_{..k} - \bar{y}_{...} \end{aligned}$$

These are "face average" sample means minus the grand average sample mean corresponding to Figure 2. It is a consequence of their definitions that $a_1 = -a_2$, $b_1 = -b_2$, and $c_1 = -c_2$.

Fitted two-factor interactions are

$$\begin{aligned} ab_{ij} &= \bar{y}_{ij.} - (\bar{y}_{...} + a_i + b_j) \\ ac_{ik} &= \bar{y}_{i.k} - (\bar{y}_{...} + a_i + c_k) \\ bc_{jk} &= \bar{y}_{.jk} - (\bar{y}_{...} + b_j + c_k) \end{aligned}$$

These are what one would call two-factor interactions in a two-way dataset after collapsing the cube in Figure 2 front-to-back, or top-to-bottom, or left-to-right by averaging. They represent what can be explained about a response mean if one thinks of factors acting jointly in pairs beyond what is explainable in terms of them acting separately. It is a consequence of their definitions that fitted 2-factor interactions add to 0 over levels of any one of the factors involved. In a 2^p study, this allows one to compute a single one of these fitted interactions of a given type and obtain the other three by simple sign changes. For example

$$ab_{11} = -ab_{12} = -ab_{21} = ab_{22}$$

(if the number of "index-switches" going from one set of indices to another is odd, the sign of the fitted effect changes, while if the number is even there is no sign change).

Finally, fitted 3-factor interactions are

$$abc_{ijk} = \bar{y}_{ijk} - (\bar{y}_{...} + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk})$$

These measure the difference between what's observed and what's explainable in terms of factors acting separately and in pairs on the response, y. They are also the difference between what one would call 2 factor interactions between, say, A and B, looking separately at the various levels of C (so that they are 0 exactly when the pattern of AB two factor interaction is the same for all levels of C). These sum to 0 over all levels of any of the factors A, B, and C, so for a 2^p factorial one may compute one of these and get the others by appropriate sign changes.

In 2^p studies, since by computing one fitted effect of each type, one has (via simple sign changes) all of the fitted effects, it is common to call those for the "all factors at their high level" combination, namely

$$a_2, b_2, ab_{22}, c_2, ac_{22}, bc_{22}, abc_{222}, \quad \text{etc.}$$

the fitted effects. And the Yates algorithm is a very efficient method of computing these "by hand" all at once. It consists of writing combinations and corresponding means down in "Yates standard order" and then doing p cycles of additions and subtractions of pairs of values, followed by division by 2^p. It is illustrated in V&J Section 4.3.5, page 189.

Having computed fitted effects for a 2^p factorial, there is a simple way of judging whether what has been computed is really anything more than background noise/experimental variation. Just as was the case for two-way factorials, each p-way factorial effect is an "L" (i.e. a linear combination of the means μ_{ijk}) and each fitted effect is the corresponding linear combination of sample means, an "$\hat{L}$" in the notation Section 21. Thus one can attach "margins of error" to fitted effects using the basic method for estimating a linear combination of means (L's) presented in that section.

The general formula from Section 21 takes a particularly simple form in the case of fitted effects from a 2^p study. Corresponding to each 2^p factorial fitted effect, $\hat{E}$ (an $\hat{L}$), is a theoretical/population effect (a corresponding L). Confidence limits for the theoretical effect are then

$$\hat{E} \pm t s_{\mathrm{P}} \frac{1}{2^p} \sqrt{\frac{1}{n_{(1)}} + \frac{1}{n_{\mathrm{a}}} + \frac{1}{n_{\mathrm{b}}} + \frac{1}{n_{\mathrm{ab}}} + \cdots}$$

In the case that the data are balanced (all samples are of size m) this formula reduces to

$$\hat{E} \pm t s_{\mathrm{P}} \sqrt{\frac{1}{m 2^p}}$$

These formulas provide the margins of error necessary to judge whether fitted effects clearly represent some non-zero real effects of the factors.

The use of confidence limits for effects requires that there be some replication somewhere in a 2^p study, so that s_{P} can be calculated. As long as someone knowledgeable is in charge of an experiment, this is not typically an onerous requirement. Getting repeat runs at a few (if not all) sets of experimental conditions is typically not as problematic as potentially leaving oneself with ambiguous results. But unfortunately there are times when a less knowledgeable person is in charge, and one must analyze data from an unreplicated 2^p study. This is a far from ideal situation and the best available analysis method is not nearly as reliable as what can be done on the basis of some replication.

All one can do when there is no replication in a 2^p factorial is to rely on the likelihood of "effect sparsity" and try to identify those effects that are clearly "bigger than noise" using normal plotting. That is,

- a kind of "Pareto principle" of effects says that in many situations there are really relatively few (and typically simple/low order) effects that really drive experimental results, and relative to the "big" effects, "most" others are small, almost looking like "noise" in comparison, and

- when effect sparsity holds, one can often identify the "few big actors" in a 2^p study by normal plotting the fitted effects, looking for those few fitted effects that "plot off the line established by the majority of the fitted effects."

Part III
Introduction to Modern (Statistical) Machine Learning

39 Some Generalities and k-Nearest Neighbor Prediction

(Text Reference/Reading: JWH&T Sections 2.1,2.2.&3.5)

The subject of most of the last three weeks of Stat 587 is an introduction to "(statistical) machine learning" or "(big data) predictive analytics" or "data mining." The James, Witten, Hastie and Tibshirani book could be a text for an entire course on this subject. In 587, we'll have time only to provide some basics that should give you perspective as to "what is here" and enable you to dig deeper on your own if the need arises.

There are two basic flavors of methodology in this field. There is **"supervised" learning**, that amounts to developing a predictor for output y based on inputs $x_1, x_2, \ldots, x_p$, and there is **"unsupervised" learning**, that concerns finding/describing patterns in $x_1, x_2, \ldots, x_p$. We'll spend our time on supervised learning.

There are then two versions of predictive analytics (supervised learning). These are

1. a "regression" version, where the y to be predicted is a genuinely quantitative (measured) variable, and
2. a "classification" version, there the y to be predicted is a category, often just 0 or 1 (or 1 and 2), but sometimes an appropriate one of $1, 2, \ldots, K$.

One begins with an $(x_1, x_2, \ldots, x_p, y)$ dataset that forms the basis for producing $\hat{y}(x_1, x_2, \ldots, x_p)$ a (hopefully good) predictor of y. In this area, this set of N cases and $p+1$ variables (giving an $N \times (p+1)$ data matrix) is usually called the set of **"training data."**

If instead of training data (almost always assumed to be iid observations from some joint distribution for $(x_1, x_2, \ldots, x_p, y)$) one had a complete understanding of the joint distribution for $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)$ and y, identification of the best possible predictor, $\hat{y}^{\text{opt}}(\boldsymbol{x})$ is relative simple.

1. In regression problems, one might seek to minimize an average squared prediction error

$$\mathrm{E}\left(y - \hat{y}(\boldsymbol{x})\right)^2$$

over choices of function $\hat{y}(\cdot)$. Some reasonably simple probability then establishes that in this case an optimal predictor is

$$\hat{y}^{\text{opt}}(\boldsymbol{x}) = \mathrm{E}\left[y|\boldsymbol{x}\right]$$

the conditional mean of y for the input $\boldsymbol{x}$.

2. In classification problems, one might seek to minimize a misclassification rate

$$P\left[y \neq \hat{y}(\boldsymbol{x})\right]$$

over choices of function $\hat{y}(\cdot)$. Some reasonably simple probability then establishes that in this case an optimal classifier (predictor) is

$$\hat{y}^{\text{opt}}(\boldsymbol{x}) = \text{the class } m \text{ maximizing } P[y = m|\boldsymbol{x}]$$

the possible value of y with the largest conditional probability for the input $\boldsymbol{x}$.

"The problem" is that one doesn't have complete knowledge of the joint distribution for $\boldsymbol{x}$ and y and can only use the training data to approximate $\hat{y}^{\text{opt}}(\boldsymbol{x})$.

So, how does one start here? We already have some ideas available from what has gone before. In a regression problem, one obvious predictor of y is the (MLR) least squares predictor

$$\hat{y}(\boldsymbol{x}) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p \tag{30}$$

for coefficients $b_0, b_1, b_2, \ldots, b_p$ derived from MLR. But in big N contexts, where many cases/instances are available in the training set and the training data are potentially adequate to support the predictor-building, we'd like to consider predictor forms much more flexible than form (30).

One very flexible simple form of prediction is built on a "**$\boldsymbol{k}$-nearest neighbor**" idea. That is the following. For prediction/classification at a vector of inputs $\boldsymbol{x}$, one finds the k cases $\boldsymbol{x}_i$ in the training set with the smallest values of

$$\text{dist}(\boldsymbol{x}, \boldsymbol{x}_i) = \sqrt{(x_1 - x_{1i})^2 + (x_2 - x_{2i})^2 + \cdots + (x_p - x_{pi})^2}$$

These are the k-nearest neighbors of $\boldsymbol{x}$ in $\Re^p$. Then considering the y_i corresponding to these k nearest neighbors, one uses

$$\hat{y}(\boldsymbol{x}) = \text{the mean } y_i \text{ for the } k \text{ nearest neighbors of } \boldsymbol{x} \text{ in the training set}$$

in regression problems, and in classifications uses

$$\begin{aligned}\hat{y}(\boldsymbol{x}) = &\text{ the class } m \text{ with the largest representation among the } y_i \\ &\text{ for the } k \text{ nearest neighbors of } \boldsymbol{x} \text{ in the training set}\end{aligned}$$

An issue with the "k-nn" idea that demands attention from a careful user is that as just described, the method is units-dependent. If one changes the units in which one of the coordinates of $\boldsymbol{x}$ is expressed, its k-nearest neighbors can change. Indeed, once one begins to notice this issue, the whole notion of computing a "distance" between input vectors whose coordinates have *any* units seems suspect. (After all, what could an expression like $\sqrt{(3\text{V})^2 + (2\text{kg})^2}$ mean?) A way of mitigating this is to operate with *standardized predictors*. That is, if coordinate l of the input vectors $\boldsymbol{x}_i$ $(i = 1, 2, \ldots, N)$ has sample mean $\bar{x}_l$ and sample standard deviation s_l, one can employ a standardized predictor $\boldsymbol{z}$ with lth coordinate

$$z_l = \frac{x_l - \bar{x}_l}{s_l}$$

These are unitless and scaled so that in every dimension the training set has standard deviation 1. "Distance" defined in terms of the standardized predictor $\boldsymbol{z}$ makes sense (and is also unitless). Rather than write "$\boldsymbol{z}$" we'll continue to write "$\boldsymbol{x}$" with the understanding that in the case of k-nn prediction the predictor has been standardized.

In predictive analytics, we must consider both the flexibility of a prediction method (its ability to produce a predictor approximating the theoretically optimal one) and the adequacy of a dataset to support its use. Often (especially when p is large) one doesn't have a large enough training set (large enough N) to make the nearest neighbor idea effective. On the other hand, MLR is often not flexible enough to approximate a (non-linear) optimal predictor. Somehow, one needs to consider a spectrum of methods of various flexibilities and find a methodology whose flexibility is as great as possible, subject to the dataset's adequacy to support its use.

A qualitative insight that can be made precise in a variety of specific situations is this:

$$\left(\begin{array}{c}\text{expected non-negative}\\ \text{penalty for imperfect}\\ \text{prediction using } \hat{y}\left(\boldsymbol{x}\right) \text{ (derived}\\ \text{from the training set)}\end{array}\right)=\left(\begin{array}{c}\text{expected non-negative}\\ \text{penalty for imperfect}\\ \text{prediction using } \hat{y}^{\text{opt}}\left(\boldsymbol{x}\right)\end{array}\right)+\left(\begin{array}{c}\text{expected non-negative penalty}\\ \text{for any difference between}\\ \hat{y}^{\text{opt}}\left(\boldsymbol{x}\right) \text{ and the best predictor}\\ \text{possible in the class considered}\end{array}\right)$$
$$+\left(\begin{array}{c}\text{expected non-negative}\\ \text{penalty for not realizing}\\ \text{the best predictor possible}\\ \text{in the class considered}\end{array}\right)$$

This says that overall "poorness" of prediction can be seen as the sum of 3 components. The first is a kind of baseline contribution that is inherent in the problem, accounting for the best one could possibly do if one knew the joint distribution of $\boldsymbol{x}$ and y. The second term is a modeling penalty, suffered because one doesn't allow enough flexibility in the choice of predictor class. (For example, if one uses a MLR predictor and $\hat{y}^{\text{opt}}\left(\boldsymbol{x}\right) = \text{E}[y|\boldsymbol{x}]$ is actually very non-linear, this term might be large. Even the best linear predictor could be much worse than $\hat{y}^{\text{opt}}\left(\boldsymbol{x}\right)$.) Finally, the last term is a fitting penalty that can be large because of randomness in the training set and/or poor technology in choosing a predictor from the class of predictors considered. Good choice of an overall prediction methodology attempts to balance the last two terms, finding a class of predictors (and effective fitting methodology) that is as flexible as possible without requiring more information than is really provided by a training set.

What can be done in practical terms toward such a goal is this. For regression problems, one can apply K-fold cross-validation as described in Section 32 to a variety of prediction methodologies and look for the one that has the smallest "prediction error"

$$\frac{1}{N}CVSSE = CVMSPE$$

and select that methodology for application to the entire training set in order to produce a final predictor $\hat{y}\left(\boldsymbol{x}\right)$. For classification problems, if $\hat{y}^{j}\left(\boldsymbol{x}\right)$ is a predictor/classifier derived from all training data except that in the jth fold, an appropriate K-fold cross-validation "prediction error rate" is

$$\frac{1}{N}\sum_{j=1}^{K}\sum_{i \text{ in fold } j}\left(\text{the number of cases } i \text{ with } y_i \neq \hat{y}^{j}\left(\boldsymbol{x}_i\right)\right)$$

One applies this measure to a variety of classification methodologies, looking for the one that has the smallest value, then applying that to the entire training set in order to produce a final classifier $\hat{y}\left(\boldsymbol{x}\right)$.

A final point in this introduction concerns the necessity of **full** cross-validation. The **entire process** of predictor development that one intends to apply after cross-validation must be applied

K times (each time with one of K folds not included in the training set) in order to reliably assess the likely effectiveness of a given method. **This includes any data preprocessing steps** like predictor standardization! For example, one cannot simply standardize once using sample means and standard deviations computed from the entire training set, but must standardize separately before building each one of the K predictors $\hat{y}^j(\boldsymbol{x})$ and finding a cross-validation error. (The point is that if one only standardizes once, since each overall sample mean $\bar{x}_l$ and sample standard deviation s_l depends upon all N cases, all N cases end up participating in the development of each $\hat{y}^j(\boldsymbol{x})$, something one is specifically *trying to avoid* in cross-validation.)

40 "Ridge," "LASSO," and "Elastic Net" Linear (Regression) Predictors

(Text Reference/Reading: JWH&T Sections 6.1&6.2)

We'll soon consider some other ways (beyond the k-nn idea) of creating highly flexible predictors from even a small number of inputs $x_1, x_2, \ldots, x_p$. But before doing that, we consider problems where one *begins* with a "large" number of inputs (p is **big**) and the difficulty to be faced is that there are usually predictors that aren't needed and there is a strong likelihood of overfitting unless one finds a way to appropriately control the flexibility of even a linear form like (30).

One idea in this direction was at least implicit in the earlier MLR material: that of "dropping some predictors" from a linear form. That is, one can consider "subset selection" for a set of predictors. Early application of this kind of thinking was phrased in terms of "all possible R^2 routines" that looked for "best" models (in terms of R^2) with a given number of predictors. (Not much formal attention was paid to consideration of avoiding overfit.) It is, of course, possible to try to compare cross-validation performance of all possible reduced models derived from a MLR full model as a more reliable way of seeking a good subset of p inputs for predicting y. But this is a very "discrete" approach to the problem of reducing predictor flexibility to a point appropriate for a given dataset, and is computationally infeasible for many modern problems (the number of reduced models to consider grows exponentially with p). Another, very clever, more continuous approach has instead gained currency.

Dropping predictors x_l from consideration can be thought of as *a priori* setting some b_l's to 0. This is in some ways enforced "shrinking" of a vector of fitted regression coefficients, $\boldsymbol{b}$, toward $\mathbf{0}$ in $\Re^p$. Modern "penalized regression" methods come at the problem of "reducing flexibility" in a linear form by a more general "shrinkage" idea. But before laying this out, we must start by noting that when *units* are involved, the numerical value (so, the "size,") of a fitted regression coefficient b_l depends upon the units used to express x_l and y. So, again as for nearest neighbor prediction, it is sensible to consider standardized predictors. If we let

$$z_l = \frac{x_l - \bar{x}_l}{s_l}$$

a fitted equation in terms of standardized inputs

$$\hat{y} = b_0 + b_1 z_1 + b_2 z_2 + \cdots + b_p z_p$$

corresponds directly to

$$\hat{y} = \left(b_0 - \left(b_1 \frac{\bar{x}_1}{s_1} + b_2 \frac{\bar{x}_2}{s_2} + \cdots + b_p \frac{\bar{x}_p}{s_p}\right)\right) + \left(\frac{b_1}{s_1}\right) x_1 + \left(\frac{b_2}{s_2}\right) x_2 + \cdots + \left(\frac{b_p}{s_p}\right) x_p$$

So coefficients for standardized predictors divided by sample standard deviations of predictors in "original units" give coefficients for predictors in original units. **We henceforth presume inputs x_l are already standardized.**

For $\lambda \geq 0$, a so-called **ridge regression** coefficient vector $\boldsymbol{b}_\lambda^{\text{ridge}}$ is a minimizer of the penalized error sum of squares

$$\text{RPen-}SSE\left(\boldsymbol{\beta}\right) = \sum_{i=1}^{N} \left(y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}\right)\right)^2 + \lambda \sum_{i=1}^{p} \beta_i^2$$

For $\lambda = 0$ this is an unpenalized error sum of squares and $\boldsymbol{b}_0^{\text{ridge}}$ is simply the ordinary least squares (standard MLR) coefficient vector. As $\lambda \to \infty$, the intercept in $\boldsymbol{b}_\lambda^{\text{ridge}}$ converges to $\bar{y}$ and all other coefficients converge to 0 (so that $\hat{y} \to \bar{y}$). For all $\lambda > 0$ the ridge penalty provides overall "shrinking" of the coefficients of the x_l (compared to their MLR counterparts) toward 0 and $\hat{y}$ toward a constant predictor. This "ridge" penalty idea thus provides a continuous spectrum of linear predictors, varying from the most flexible (for $\lambda = 0$) to the least (for very large λ). One can hope to compare them by cross-validation and find a λ and corresponding $\hat{y}$ that provides good predictions.

A first variant of this idea is so-called "**lasso**" (least absolute shrinkage and selection operator) penalized **regression**. For $\lambda \geq 0$, a so-called **lasso** coefficient vector $\boldsymbol{b}_\lambda^{\text{lasso}}$ is a minimizer of the penalized error sum of squares

$$\text{LPen-}SSE\,(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})\right)^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

The penalty here involves absolute values of coefficients rather than their squares. Although it is probably not obvious to the reader, this change has the effect of tending to make some individual entries of $\boldsymbol{b}_\lambda^{\text{lasso}}$ exactly 0, i.e. of exactly "zeroing out" dependencies of $\hat{y}$ on some of the x_l, and thereby accomplishing a kind of automatic or continuous "variables selection."

For $\lambda = 0$ the lasso-penalized error sum of squares is again the unpenalized error sum of squares and $\boldsymbol{b}_0^{\text{lasso}}$ is once more simply the ordinary least squares (standard MLR) coefficient vector. As $\lambda \to \infty$, the intercept in $\boldsymbol{b}_\lambda^{\text{lasso}}$ converges to $\bar{y}$ and all other coefficients converge to 0 (so that $\hat{y} \to \bar{y}$). As in ridge regression, for all $\lambda > 0$ the lasso penalty provides overall "shrinking" of the coefficients of the x_l (compared to their MLR counterparts) toward 0 and predictors toward the constant predictor $\hat{y} = \bar{y}$. The number of individual coefficients in $\boldsymbol{b}_\lambda^{\text{lasso}}$ that are exactly 0 tends to increase with λ. (It is *not the case*, however, that because a β_l is zeroed out at a particular value of λ it must necessarily remain zeroed out for larger λ.) Like ridge penalization, the lasso idea thus provides a continuous spectrum of linear predictors, varying from the most flexible (for $\lambda = 0$) to the least (for very large λ). One can hope to compare them by cross-validation and find a λ that provides good predictions.

A generalization of both ridge and lasso ideas is so-called "**elastic net**" penalized **regression**. For constants $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ a so-called elastic net coefficient vector $\boldsymbol{b}_\lambda^{\text{enet}}$ is a minimizer of the penalized error sum of squares

$$\text{ENPen-}SSE\,(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi})\right)^2 + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p} \beta_i^2$$

The penalty here involves both ridge- and lasso-type parts. The lasso part of the penalty tends to provide some effect of exactly "zeroing out" dependencies of $\hat{y}$ on some of the x_l, as in lasso regression accomplishing a kind of automatic variables selection. Since by choosing one of the coefficients λ to be 0 one gets both ridge and lasso penalties as special cases, *the best elastic net regression predictor is at least as good as the best ridge or the best lasso predictor* (either one of which is by the same logic at least as good as the ordinary least squares MLR predictor!).

For $\lambda_1 = \lambda_2 = 0$ the elastic net penalized error sum of squares is the unpenalized error sum of squares and $\boldsymbol{b}_{\mathbf{0}}^{\text{enet}}$ is simply the ordinary least squares (standard MLR) coefficient vector. As one or both of the λ's grow the intercept in $\boldsymbol{b}_{\boldsymbol{\lambda}}^{\text{enet}}$ converges to $\bar{y}$ and all other coefficients converge to 0 (so that $\hat{y} \rightarrow \bar{y}$). For at least one $\lambda > 0$, the elastic net penalty provides overall "shrinking" of the coefficients of the x_l (compared to their MLR counterparts) toward 0 and predictors toward the constant predictor $\hat{y} = \bar{y}$. The elastic net idea thus provides a *doubly indexed* continuous spectrum of linear predictors, varying from the most flexible (for $\lambda_1 = \lambda_2 = 0$) to the least (for very large λ_1 or λ_2). One can hope to compare them by cross-validation and find a pair (λ_1, λ_2) that provides good predictions.

The fitting of ridge, lasso, and elastic net regressions in `R` is effectively done using the `glmnet` package in `R`. The `caret` package in `R` is also an extremely useful one, in that it makes doing cross-validation and corresponding comparison of multiple parameter sets for various prediction packages (including `glmnet`) quite routine by providing one standard input and function call format. It has an option (that surely should be used with the methods of this section) for automatically handling the redoing of standardization for each fold of a cross-validation study.

Ultimately, the use of the methods of this section help one avoid overfitting in a "large p" linear prediction problem by more or less damping the effects of inputs on predictions (over what would be suggested by least squares MLR). Fitted predictions are shrunken towards $\bar{y}$ and made less flexible than what least squares alone would prescribe, in a way somewhat like what is prescribed by consideration of reduced models in "ordinary" MLR.

41 Tree Predictors (Regression and Classification Trees)

(Text Reference/Reading: JWH&T Section 8.1)

A flexible prediction methodology that is unlike anything we have discussed thus far is one based on (binary) "trees." These are predictors (of both regression and classification types) that are constant on "rectangular regions" of input space $\Re^p$ defined by sequentially splitting an existing rectangular region into two parts on the basis of whether a particular (well-chosen) input variable (say, x_l) is larger or smaller than a particular (well-chosen) value. This kind of development of a set of regions is effectively summarized graphically with a binary tree. Figure shows a simple $p = 2$ example made using 4 splits and therefore having 5 "rectangular regions" indicated as R1, R2, R3, R4, and R5. A predictor based on this tree would be constant on each of the 5 regions.

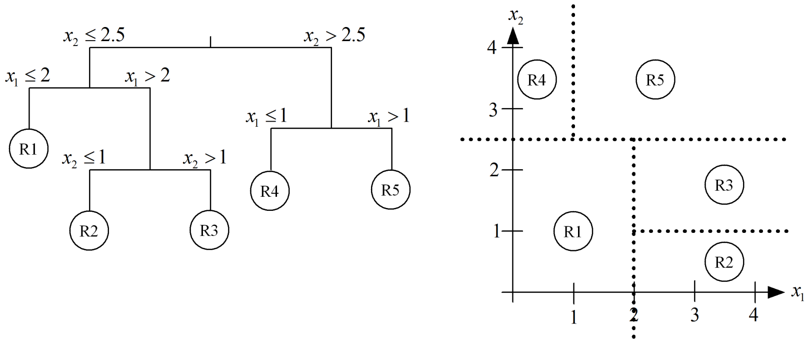

Figure 3: Hypothetical Binary Tree for a $p = 2$ Case

For the time being putting aside the question of how one *looks for* an appropriate tree structure, consider the issue of what training-set-based predictor to use for a given tree. In the "**regression" case** one is going to make $\hat{y}$ constant on each region R and judge training set performance in terms of sum of squared prediction errors. So if $\mathrm{R}(\boldsymbol{x})$ is the region containing $\boldsymbol{x}$, it is clear that a best choice of constant prediction for the region is simply

$$\hat{y}^{\text{tree}}(\boldsymbol{x}) = \bar{y}_{\mathrm{R}(\boldsymbol{x})} = \text{sample mean response for those training cases with } \boldsymbol{x}_i \text{ in the same region } (\mathrm{R}(\boldsymbol{x})) \text{ as } \boldsymbol{x}$$

In the **classification case**, one is going to make $\hat{y}$ constant on each region R and judge training set performance in terms of training set misclassification rate. So it is clear that a best choice of constant prediction for the region is simply

$$\hat{y}^{\text{tree}}(\boldsymbol{x}) = \text{the most frequent class (value of } y_i\text{) for those training cases with } \boldsymbol{x}_i \text{ in the same region } (\mathrm{R}(\boldsymbol{x})) \text{ as } \boldsymbol{x}$$

So now consider the problem of finding a "good" binary tree structure to use in prediction. What is possible and common is a two-step process consisting of

1. "growing" a large tree using a "forward selection" heuristic guided by a measure of overall consistency of training set y_i's for the $\boldsymbol{x}_i$'s in each region, (usually) followed by

2. a "pruning" step where optimal sub-trees of the large tree from 1. are identified and naturally linked with values of a penalty/complexity parameter.

(The complexity parameter is chosen by cross-validation and an optimal sub-tree of the heuristically derived large tree based on the whole training set is then employed for final prediction.)

To provide some more details, consider first the building of the large tree referred to in 1. For $m \geq 0$ suppose that m splits have been made, making $m+1$ rectangular regions. In order to choose an $(m+1)$st split, one considers all splits of all possible existing rectangular regions between all values possible of all p variables x_l, and chooses the one providing the greatest reduction in some measure of "impurity" of the y_i's for $\boldsymbol{x}_i$'s within each region. For **regression cases**, the most natural impurity measure is the **error sum of squares** for $\hat{y}^{\text{tree}}(\boldsymbol{x})$ (one looks for a split of one rectangle providing the largest decrease in SSE). For **classification cases**, the most natural impurity measure[1] is the **empirical misclassification rate (the training set error rate)** for $\hat{y}^{\text{tree}}(\boldsymbol{x})$ (one looks for a split providing the largest increase in training set classification accuracy). (In both instances, if rectangles are completely homogeneous with respect to their associated values of y_i, one has an apparently perfect predictor.) The tree-building process continues until a "perfect" predictor is produced or some preset upper limit for number of rectangles is reached.

In order to describe "optimal" pruning of a tree, we must first consider a complexity-penalized figure of merit to associate with a tree predictor. For a tree T and associated predictor $\hat{y}^{\text{tree}}(\boldsymbol{x})$, let $E(T)$ stand for

1. $\dfrac{SSE}{N}$ in a regression problem, or

2. the empirical (training set) misclassification rate in a classification case.

Then for $|T|$ the number of rectangular regions defining the predictor (the number of splits made plus 1) and $\alpha > 0$, a penalized prediction cost to associate with T is

$$C_T(\alpha) = E(T) + \alpha |T|$$

For a large binary tree S grown for prediction (regression or classification) from a training set, it turns out to be possible to efficiently consider all possible sub-trees of S that can be produced by "cutting" the tree diagram at any subset of its branch points[2], and find the one minimizing $C_S(\alpha)$ for each α. Call the best sub-tree for the value α by the name S_α. As α increases, the size of the optimal sub-tree, $|S_\alpha|$, decreases. The $\alpha = 0$ case is the full tree, S. The very large α case is that where no splits are made. It is worth noting that as α goes from 0 to ∞, it will typically *not* be the case that $|S_\alpha|$ takes every integer value from $|S|$ to 0. There will typically be some potential sizes of sub-trees for which no sub-tree of that size is best *for any value of* α.

Tree predictors are very flexible and (at least when they are based on relatively few splits) are easy to interpret. They also have the interesting feature that they are completely scale-independent as regards the inputs, x_l. Since transforming any of the inputs using an increasing or decreasing transformation doesn't change the predictor, using a tree predictor eliminates the question of how to transform individual inputs.

[1] There are other impurity measures sometimes recommended for classification problems. Two are the so-called "Gini index" and the so-called "cross entropy."

[2] There are many more sub-trees possible than just those met in the building of S.

42 Bootstrapping, Bagging, and Random Forests

(Text Reference/Reading: JWH &T Section 8.2.1 & 8.2.2)

A way of making "new" datasets that "look like" a training set *without looking "too much" like the training set* is known as **bootstrapping**. A bootstrap sample from a training set of size N is a dataset made by sampling N cases at random *with replacement* from the the training set. The result is a random variant of the training set that (as long as N is at all large, say at least 20) is virtually guaranteed to miss some cases in training set and include multiple copies of others. (It's fairly easy to argue that on average about 37% of the training set will be missed in the bootstrap sample for such N). A predictor built from a bootstrap sample is then a random variant of a predictor built from the original training set.

The idea of "**bagging**" (so-called *bootstrap aggregation*) is to make many (say B) bootstrap samples, create a predictor from each, and to "aggregate" these into a single predictor. The hope is to reduce the tendency to overfit (by virtue of not using a given case in making a fair number of the constituent predictors) while nevertheless remaining true to overall patterns in the training set (by virtue of making only random variants of the full training set as bootstrap samples).

To be somewhat more precise, for bootstrap sample b, let $\hat{y}^b(\boldsymbol{x})$ be the corresponding predictor built using a particular methodology of interest. (One can "bag" *any* form of predictor, though the discussion in JWH&T Section 8.2 reads as if the methodology concerns only tree predictors.) Then in regression contexts, a bagged version of the predictor is

$$\hat{y}^{\text{bagged}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B}\hat{y}^b(\boldsymbol{x})$$

the sample mean of the B predictors built from the bootstrap samples. For classification contexts, the obvious way to aggregate is to set

$$\begin{aligned}\hat{y}^{\text{bagged}}(\boldsymbol{x}) = {} & \text{the most frequent class (value of } \hat{y}^b(\boldsymbol{x})\text{) for} \\ & \text{the } B \text{ predictors built on bootstrap samples}\end{aligned}$$

One expects that for large B the predictions and corresponding performance of $\hat{y}^{\text{bagged}}(\boldsymbol{x})$ converge to those of some limiting case (as the effect of random selection of the bootstrap samples is averaged out across a large number of "trials").

In an important and happy development, bagging more or less *provides its own cross-validation.* That is, since any particular training case i is missed by about 37% of the B bootstrap samples, if its y_i is predicted *using those bootstrap samples only*, one might hope to accurately approximate prediction performance of the methodology. That is, let $\hat{y}_i^*$ be a version of the bagged prediction at $\boldsymbol{x}_i$ computed using only those bootstrap samples that do not include case i. In regression contexts a so-called **"out-of-bag" mean square prediction error** is

$$OOB = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i^*)^2$$

and in classification problems an **out-of-bag classification error rate** is

$$OOB = \frac{1}{N}\,(\text{the number of } y_i \neq \hat{y}_i^*)$$

These depend upon B but converge with increasing B to a number that can be expected to represent the real prediction performance of the method being considered. In practice, one plots these against B looking for when the plot levels off as a means of judging when B is large enough for use in a particular application.

A very important application of bagging is the so-called **"random forest"** predictor. This is bagging of a specialized kind of tree. A random forest is made by operating on each of B bootstrap samples to produce a constituent tree as follows. All nodes are split until any further split of a node would produce one with fewer than "*nodesize*" cases in it. In splitting a particular node, some number, m, of the p dimensions of the input vector $\boldsymbol{x}$ are chosen at random for consideration, and the best split possible for (only) those coordinates x_l of $\boldsymbol{x}$ is selected. The parameters m and *nodesize* can be optimized for the best performance in terms of "large B" values of OOB. When this is not done, more or less standard default values for parameters are

1. $m = \lfloor p/3 \rfloor$ and $nodesize = 5$ in regression problems, and

2. $m = \lfloor \sqrt{p} \rfloor$ and $nodesize = 1$ in classification problems.

A different kind of control on how complex a particular tree in the forest can become that is sometimes used with or in place of the *nodesize* constraint, is limitation on the depth of any final node of the tree. (For example, one might split nodes until every final node is no more than $depth = 4$ splits from the top of the tree.) All of *nodesize*, *depth*, and m are complexity parameters that can be optimized (in terms of limiting OOB) in choice of a good random forest.

Random forests have been remarkably effective in a wide variety of applications. The admittedly rather odd-sounding rules by which they are produced often turn out to produce very good predictors.

43 Boosting and Stacking (Trees and Other Regression Predictors)

(Text Reference/Reading: JWH&T Section 8.2.3)

Bagging is an "ensemble" prediction method, in that it combines a number of different predictors into a single overall predictor. The ensemble notion has several other implementations in modern predictive analytics. Two that we will introduce here are known as "boosting" and "stacking."

Boosting amounts to successive correction of a current predictor by small perturbations intended to (successively) improve prediction (essentially by broadening the form allowed for an ultimate predictor). Consider the "regression" prediction problem, where the response variable, y, is quantitative. Take an initial predictor to be

$$\hat{y}_0(\boldsymbol{x}) = \bar{y}$$

Then with an mth version of a predictor, say $\hat{y}_m(\boldsymbol{x})$ in hand, make mth "residuals"

$$e_i^m = y_i - \hat{y}_m(\boldsymbol{x}_i)$$

Choose/fit some good predictor of "responses" e_i^m, say $\hat{e}^m(\boldsymbol{x})$. Then for some (typically small) factor $0 < \nu < 1$, create an $(m+1)$st predictor as

$$\hat{y}_{m+1}(\boldsymbol{x}) = \hat{y}_m(\boldsymbol{x}) + \nu \cdot \hat{e}^m(\boldsymbol{x})$$

This is $\hat{y}_m(\boldsymbol{x})$ corrected for the fraction ν of its failure to perfectly predict y. Complexity parameters here are

$$M = \text{the number of boosting iterations for the final predictor}$$

and ν, predictor complexity increasing with both. Rational choice of them to produce

$$\hat{y}^{\text{boost}}(\boldsymbol{x}) = \hat{y}_M(\boldsymbol{x})$$

proceeds by cross-validation.

The choice of the form of predictors for residuals in boosting is not limited to any one of the methods considered here, but by far the most common implementation employs trees. The end result of doing (regression-type) boosting with trees is to make a predictor that is a linear combination of trees. (In this context, another complexity parameter will be the depth to which one is allowed to build a tree $\hat{e}^m(\boldsymbol{x})$.) A currently extremely popular form of boosting with trees is the so-called "`XGBoost`" (extreme gradient boosting) algorithm that has a very fast/effective implementation in `R` (and other languages like `python` commonly used in machine learning).

There are versions of boosting appropriate to classification problems, the most famous of which is the "`AdaBoost.M1`" algorithm. Recently, a version of `XGBoost` appropriate to classification has gotten more attention than the original `AdaBoost.M1`, probably mostly because of the former's superior implementation. Exact forms of boosting for classification are harder to motivate and describe than boosting for regression problems, so we will not try to present them here.

The fact that a boosted regression predictor is "a linear combination of trees" suggests the possibility of simply beginning with several regression predictors, say $\hat{y}_1(\boldsymbol{x}), \hat{y}_2(\boldsymbol{x}), \ldots, \hat{y}_M(\boldsymbol{x})$, corresponding constants $c_1, c_2, \ldots, c_M$, and then using an ensemble regression predictor

$$\hat{y}^{\text{stacked}}(\boldsymbol{x}) = \sum_{m=1}^{M} c_m \hat{y}_m(\boldsymbol{x}) \tag{31}$$

that is a linear combination of the M separately-derived predictors. The obvious fact that each $\hat{y}_m(\boldsymbol{x})$ can be obtained by choosing $c_m = 1$ and all other coefficients 0 implies that one can always do at least as well using an effectively chosen stacked predictor as can be done using any single element of the ensemble. The "trick," of course, is effective choice of the constants c_m.

The linear form (31) suggests a variety of "generalized stacking" methodologies. One can essentially think of the first-level predictors $\hat{y}_1(\boldsymbol{x}), \hat{y}_2(\boldsymbol{x}), \ldots, \hat{y}_M(\boldsymbol{x})$ as a set of clever "features" to be treated as inputs to *any* final prediction methodology of interest. Due to their scale-independence, tree-based methods (ordinary trees or random forests or tree-based boosting) are particularly attractive at the top level— arguably more so than ordinary (linear-combination-type) stacking. The choice of method complexity parameter(s) is then the fundamental problem to be overcome. As always, cross-validation is the most effective guide to the simultaneous choice of the parameter(s) of each of the first-level predictors $\hat{y}_1(\boldsymbol{x}), \hat{y}_2(\boldsymbol{x}), \ldots, \hat{y}_M(\boldsymbol{x})$ and the top-level methodology. The main obstacle to success is the huge amount of computing implied by its use.

There are more or less obvious versions of "generalized stacking" that can be applied to classification problems. For $\hat{y}_1(\boldsymbol{x}), \hat{y}_2(\boldsymbol{x}), \ldots, \hat{y}_M(\boldsymbol{x})$ classifiers or underlying assessments of $P[y = 1|\boldsymbol{x}]$, treating them as features entered into a fixed final classification methodology is a way to make an effective final ensemble classifier. The practical limitation is again the feasibility of cross-validation to guide choice of tuning parameters and assess likely final performance.

The main attraction of ensemble predictors is the fact that they offer flexibility/complexity beyond that provided by any single method used in isolation. Their main limitation is the difficulty of control of that flexibility through proper cross-validation that does with K training sets (each missing a fold of the training set) what will ultimately be done with the entire training set to make an ensemble predictor.

44 Smoothing and Generalized Additive Model (Regression) Prediction

(Text Reference/Reading: JWH&T Sections 7.6-7.7)

The issue of "feature engineering" in prediction can be thought of as the problem of potentially replacing (or supplementing) individual predictors x_l (or several predictors x_l) with transformations (functions) of them that work better as inputs to standard predictor forms than the x_l themselves. Of course, identifying *what* features will be effective in a given problem is "the difficulty." **Smoothing methods** and their use in "generalized additive" prediction can to some degree be thought of as an "automatic" means of feature selection. Rather than more or less rummaging through one's repertoire of familiar useful functions (logs, trig functions, exponentials, etc.) smoothing attempts to custom-build new predictors from existing ones[3].

There are two main lines of development of smoothing methods. So-called **smoothing splines** are introduced in Section 7.5 of JWH&T and (local averaging and) **local regression methods** are introduced in their Section 7.6. The latter are easier to describe than the former, and since time is short and often the two approaches give similar results, here we will consider only local (averaging and) regression methods and their use in generalized additive modeling (treated in Section 7.7 of JWH&T).

Temporarily suppose that only a single predictor, x, is under consideration ($p = 1$). For a basic "smoothing kernel" $D(t) \geq 0$ that is symmetric in t around 0 and decreases in $|t|$ (often chosen so that $D(t) = 0$ for $|t| > 1$) we first consider building weighted average predictors based on $D(t)$. For concreteness sake, one can consider the choice

$$D(t) = \phi(t)$$

where the basic kernel is the standard normal pdf. Three standard choices of kernels are portrayed in Figure 4, where the choice $D(t) = \phi(t)$ is plotted in red.

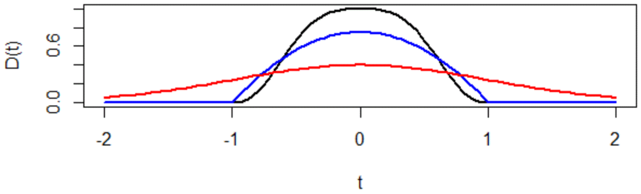

Figure 4: Three Standard Smoothing Kernels $D(t)$. Normal is in Red. "Epanechnikov" is in Blue. "Tricube" is in Black.

For $\lambda > 0$ a "bandwidth" and prediction at a value x, we consider weighting the values $x_1, x_2, \ldots, x_N$ through $D(t)$ and their distances to x as

$$w_\lambda(x_i, x) = D\left(\frac{x_i - x}{\lambda}\right)$$

[3] Tree-based prediction methods in some sense avoid this issue altogther in that making increasing or decreasing transformations of input variables has no effect on tree predictions.

(The larger the bandwidth, the more slowly weights decay as one moves away from x.)

Then a weighted average predictor (the so-called Nadarya-Watson predictor) is

$$\hat{y}_{\lambda}^{\mathrm{NW}}(x)=\frac{\sum_{i=1}^{N} y_i w_{\lambda}(x_i, x)}{\sum_{i=1}^{N} w_{\lambda}(x_i, x)}$$

This predictor is a smooth function of x. For large λ it is essentially $\bar{y}$. For small λ it is very much controlled by the value of y at the training case closest to x. λ is a complexity parameter that can be chosen by cross-validation.

Even for wisely chosen values of the bandwidth, $\hat{y}_{\lambda}^{\mathrm{NW}}(x)$ has deficiencies related to its inability to track a trend in y at the left or right ends (regarding x) of a training set, and at places "inside" the dataset where the values x_i are relatively sparse. Two modifications of the basic "local averaging" notion address this problem. These are 1) the replacement of local averaging with local regression and 2) making the bandwidth "adaptive."

The (simple linear)[4] local regression idea is this. For prediction at x, one uses a value from a "local regression line"

$$\hat{y}_{\lambda}^{\mathrm{local}}(x)=b_0^{\lambda}(x)+b_1^{\lambda}(x)\cdot x$$

whose coefficients $b_0^{\lambda}(x)$ and $b_1^{\lambda}(x)$ minimize the *locally weighted* error sum of squares

$$SSE_{\lambda}^{x}(b_0, b_1)=\sum_{i=1}^{N} w_{\lambda}(x_i, x)(y_i-(b_0+b_1 x_i))^2$$

(This down-weights the impact of poor predictions far from x in the choosing of coefficients b_0 and b_1 for the line used to predict y at x.)

The idea of making the bandwidth λ adaptive is to replace direct choice of a single λ with choice of a parameter "*span*" that governs roughly how much of a training set is involved in the choice of local regression coefficients at x. For example, if the parameter *span* is set to .25, at each different x a bandwidth is chosen so that only the $.25N$ training cases with x_i closest to x get any appreciable weight $w_{\lambda}(x_i, x)$, and thus have any influence on the choice of coefficients b_0 and b_1. So, instead of choosing a single bandwidth via cross-validation, one chooses a single *span* via cross-validation as a means of identifying a level of smoothing that the training set will support.

Moving the local regression idea beyond the case of a single predictor to the case where $\boldsymbol{x}\in\Re^p$ is at least in theory completely straightforward. For prediction at x, one uses a "local MLR"

$$\hat{y}_{\lambda}^{\mathrm{local}}(\boldsymbol{x})=b_0^{\lambda}(\boldsymbol{x})+b_1^{\lambda}(\boldsymbol{x})\cdot x_1+b_2^{\lambda}(\boldsymbol{x})\cdot x_2+\cdots+b_p^{\lambda}(\boldsymbol{x})\cdot x_p$$

whose coefficients $b_0^{\lambda}(\boldsymbol{x}), b_1^{\lambda}(\boldsymbol{x}), \ldots, b_p^{\lambda}(\boldsymbol{x})$ for weights

$$w_{\lambda}(\boldsymbol{x}_i, \boldsymbol{x})=D\left(\frac{\mathrm{dist}(\boldsymbol{x}_i, \boldsymbol{x})}{\lambda}\right)$$

("dist" meaning $\Re^p$ distance) minimize the weighted error sum of squares

$$SSE_{\lambda}^{x}(b_0, b_1, \ldots, b_p)=\sum_{i=1}^{N} w_{\lambda}(\boldsymbol{x}_i, \boldsymbol{x})(y_i-(b_0+b_1 x_{1i}+\cdots+b_p x_{pi}))^2$$

[4] More complicated versions of this idea can do weighted quadratic or higher order polynomial regression at x.

(This down-weights the impact of poor predictions far from $\boldsymbol{x}$ in the choosing of coefficients $b_0, b_1, \ldots, b_p$ for the linear form used to predict y at $\boldsymbol{x}$.) And the *span* idea translates directly to local MLR using regular Euclidean distance in $\Re^p$.

It is operationally straightforward to do local regression smoothing in high dimensions (for large p). But in practical terms, for p much larger than 2 or 3 local regression typically suffers from the same kind of tendency to overfit as does k-nn prediction. When smoothing is to be helpful in high dimensions, it needs to applied to a few variables at a time. This is the motivation behind **generalized additive** modeling/prediction.

To fix ideas, consider a $p = 3$ case with predictor $\boldsymbol{x} = (x_1, x_2, x_3)$. A generalization of the basic MLR form

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_2$$

is a form

$$\hat{y} = g_1(x_1) + g_2(x_2) + g_3(x_3)$$

for arbitrary smooth functions g_1, g_2, and g_3. It turns out to be possible to use the local regression smoothing ideas (1 dimension at a time, iteratively until convergence) to empirically make approximations of best versions of these functions, say $\hat{g}_1, \hat{g}_2$, and $\hat{g}_3$. Operationally, this is accomplished using a "generalized additive model" fitting program in `R`. The resulting predictor

$$\hat{y} = \hat{g}_1(x_1) + \hat{g}_2(x_2) + \hat{g}_3(x_3)$$

amounts to a kind of continuous input "(arbitrary smooth) main effects only" predictor. It is even possible to use the technology to fit a form like

$$\hat{y} = g_1(x_1) + g_2(x_2) + g_3(x_3) + g_4(x_1, x_2)$$

that includes continuous input two-factor interactions of x_1 and x_2. And so on. As the number of terms used as arguments of a g increases beyond 1, the size of the training set needed to make the technology effective skyrockets. But this idea does provide a way forward for high-dimensional flexible feature extraction.