Political Science Math Camp 2015 (Second Half)

Jason S. Davis

August 31-September 4, 2015

Contents

1 Welcome/Introduction

- Jason Davis - E-mail: jasonsd@umich.edu - Office: Haven Hall 7730 - Telephone: (write in class)
- Friday class schedule will be different: will meet for three hours in the afternoon, of which the first hour will be lecture and the last two will be the final test.
- Second half is more oriented towards preparing you for 599/the statistics sequence generally, but you'll find that much of the content is applicable to both statistical methods and formal theory.
- Donuts are Dimo's Deli and Donuts! Best donuts in Ann Arbor.

2 Basic building blocks

2.1 Logic

- Will sometimes use logical notation. I.E. $A \to B$, $A \to \neg B$, $A \leftrightarrow B$. I don't intend to spend too much time on logic in this course, but it's useful for exposition.
- $A \to B$. What if we have $\neg B$? What does this imply about A?
- What if we have B? Do we know anything about A? Fallacy of affirming the consequent.
- AND: both elements must be true, represented $\wedge$. OR: on element must be true, represented $\vee$.
- Say we have $A \vee B$. We have A. What do we know about B?
- Say we have $\neg B$. What do we know about the statement $A \wedge B$?
- Proofs are based on using a series of logic statements to show that some B is implies by A. Formal models are thus just this: series of logical statements.
- Conditionals versus biconditions.
- Review of some laws:
  - DeMorgan's Laws: $\neg(P \wedge Q) \leftrightarrow \neg P \vee \neg Q$ and $\neg(P \vee Q) \leftrightarrow \neg P \wedge \neg Q$
  - Distributive $P \vee (Q \wedge R) \leftrightarrow (P \vee Q) \wedge (P \vee R)$
  - $\neg\neg P \leftrightarrow P$
- Example: Simplify $P \vee (Q \wedge \neg P)$
- Composite statements. e.g. $A \to (B \to C)$
- Quantifiers. $\forall x,\ P(x)$ or $\exists x,\ s.t.\ P(x)$
- May need to convert natural language statement into formal logical structure.
- Everyone in the class thinks Jason is great: $\forall x \in C, P(x)$ where $P(x)$ represents "Jason is great"
- Equivalent formulation: $\nexists x \in C\ s.t. \neg P(X)$

3 Introduction to Set Theory, Probability, and Statistics

- Sample space is the set of all possible outcomes. Denote this S.
- An "event" is some subset of outcomes. Denote this $A \subseteq S$
- What is the sample space of rolling a die? How about rolling two dice? Write the ordered pairs $S = \{(1,1),(1,2),...,(6,6)\}$.
- Say we wanted to define the event of rolling a six. What subset $A \subseteq S$ represents this event?

3.1 Algebra of Sets: Intersections and Unions

- Definition: The complement of a set A (denoted A^c) is the set of all elements of S that do not belong to A.
- In terms of events, this is when event A did not happen. Return to one die - what is the complement of rolling a six?
- Definition: The intersection of A and B, denoted $A \cap B$, is the set of all elements that belong to *both* A and B.
- Definition: The union of A and B, denoted $A \cup B$, is the set of all elements that belong to *either* A or B.
- Return to two dice. Let A be the event that the two dice add to 5. Let B be the event that both die are even numbers. What is $A \cup B$? What is $A \cap B$?
- Disjoint/mutually exclusive iff $A \cap B = \emptyset$
- Unions of multiple sets $\bigcup_{i=1}^{n} A_i$, intersections of multiple sets $\bigcap_{i=1}^{n} A_i$.
- Similar to summation notation ($\sum_{i=1}^{4} x = x + x + x + x$. Can generalize using i index.) and product operator ($\Pi_{i=1}^{4} x = x^4$).
- Examples with dice.
- Countable versus uncountable sample spaces? To be continued!

3.2 Introduction to Probability

- Look to define a probability function that assigns probabilities to events.
- Axiom 1: Let A be any event defined over S. Then $P(A) \geq 0$.
- Axiom 2: $P(S) = 1$
- Axiom 3: Let A and B be any two mutually exclusive events defined over S. Then $P(A \cup B) = P(A) + P(B)$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B) = P(A) + P(B) - P(A \cup B)$
- What does it mean to say that two variables are independent?
- Answer: $P(A \cap B) = P(A)P(B)$

- If they are not independent, then the probability of one depends on whether the other occurs or not. Simplest example involves mutually exclusive events, e.g. $P(B|A) = 0$.
- People sometimes confuse mutual exclusivity and independence, when in fact they couldn't be more different (they are themselves mutually exclusive).
- Bayes Rule used to find conditional probabilities.

3.3 Conditional probability

- Conditional probability is the probability of some event A *given* that some other event B has already occurred.
- This has the effect of shrinking the sample space.
- Consider a simple example with two die. What's the unconditional probability of rolling a twelve? What is the conditional probability of rolling a twelve, given that your first roll returned a six?
- Written as $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- This can also be written $P(A|B) = \frac{P(A \cap B)}{P(B \cap A) + P(B \cap A^C)}$
- Conditional probability can sometimes be counterintuitive. Consider the Monty Hall problem (from the show *Let's Make a Deal*).
- Prize is behind randomly selected door of three doors. Monty Hall (who, incidentally, was born in Winnipeg, Manitoba, Canada) would then open one of the other doors to show that there was no prize behind it. The contestant would then be offered the opportunity to switch doors.
- Consider an example case where you choose Door 2. At the point of choosing, you have a one-third chance of it being the correct door, and each other door has a one-third chance
- Then Monty (that knave) opens a door. Say it's door 3 (opening door 1 would produce a symmetric problem), and note that Monty randomizes opening doors, with the exception that Monty never opens the door with the prize, and never opens the door you picked. Should you switch doors (i.e. switch to choose door 1)?
- Phrased differently, what's the probability it's behind each door?
- Probability it's behind door 1 is now 2/3! So yes you should switch. We'll find out why in a second.
- Use Bayes' Theorem for calculating conditional probabilities.
- $P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\neg A)P(\neg A)} = \frac{P(B|A_j)P(A_j)}{\Sigma_i P(B|A_i)P(A_i)}$
- Law of total probability: $P(A) = \Sigma_i P(A|B_i)P(B_i)$
- Practice example: Say that 1% of the population has cancer. Now say you have a test for cancer that correctly states you have cancer 99% of the time if you have cancer, but returns a false positive 1% of the time even if you don't. You take the test and it says you have cancer. What is the probability that you have cancer?
- $P(cancer|+) = \frac{P(+|cancer)P(cancer)}{P(+|cancer)P(cancer) + P(+|\neg cancer)P(\neg cancer)} = \frac{(1)(0.01)}{(1)(0.01) + (0.01)(0.99)} = 0.5$
- Intuition: Say you randomly selection 100 people from the population. On average, approximately one will have cancer and test positive, and one will not have cancer and test positive. So if you are tested positive, you have a 50/50 chance of being either.

- So back to Monty Hall problem. What we want to compare is $Pr(\textit{behind door } 1|\textit{opened door } 3)$ and $Pr(\textit{behind door } 2|\textit{opened door } 3)$.
- $Pr(BD1|OD3) = \frac{Pr(OD3|BD1)Pr(BD1)}{Pr(OD3|BD1)Pr(BD1)+Pr(OD3|BD2)Pr(BD2)+Pr(OD3|BD3)Pr(BD3)}$ $= \frac{(1)(1/3)}{(1)(1/3)+(0.5)(1/3)+(0)(1/3)} = \frac{2/6}{2/6+1/6+0} = \frac{2}{3}$.
- $Pr(BD2|OD3) = \frac{Pr(OD3|BD2)Pr(BD2)}{Pr(OD3|BD2)Pr(BD2)+Pr(OD3|BD1)Pr(BD1)+Pr(OD3|BD3)Pr(BD3)}$ $= \frac{(0.5)(1/3)}{3/6} = \frac{1}{3}$
- Intuition here: There was initially a one third probability it's behind the door you picked, and 2/3 probability it's behind a door you didn't pick. When Monty picks a door, if it's behind the door you picked then Monty can randomize, but if it's behind a door that you didn't pick, Monty *has* to pick the door which doesn't have it. So that 2/3 probability weight gets transferred to the one remaining door.
- Independence: $P(A|B) = P(B)$.

3.4 Combinatorics

- Combinatorics, radically simplified, is a branch of mathematics that looks at groups of elements from finite sets.
- In general, we use combinatorics to determine how many different ways elements can be combined.
- This ends up being super useful in probability theory, and in social sciences generally.
- Factorials: e.g. 5!. Compute by $5 * 4 * 3 * 2 * 1$. This gives you how many different ways you can organize five elements in groups of five where order matters. General form of $n!$.
- Weird case: $0! = 1$? There's a math reason for this I don't remember. Maybe Jean knows?
- How to compute factorials when you're dividing them. $\frac{8!}{4!} = 8 * 7 * 6 * 5$. Lots of stuff cancels out.
- Order might matter in, for instance, cases where you are rearranging letters. Example: arrange $\{A, B, C, D, E\}$.
- Permutations: $P(n, k) = \frac{n!}{(n-k)!}$. This allows you to compute how many different ways you can arrange n elements in groups of k where order matters.
- Note that $n!$ from above is a special case: if you arrange n elements in groups of n you get $\frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$
- Why is this (weakly) smaller than when you arrange all the elements? Consider that if you have five elements, $\{B, A, R, D, E\}$. If you are arranging in groups of three elements, and example of a permutation would be BAR. But for every such permutation, you would be able to generate more groups if you were combining more elements: with all five, you could create $\{BAR-DE, BAR-ED\}$. So the permutations formula "divides out" some of the groups.
- Permutations can be useful for probability: if we randomly select three letters (in sequence) from a bag containing all five of the letters above, what's the probability we get the word "bar"?
- **Ans:** There is one combination of three letters that gets us to $B - A - R$, and $\frac{5!}{(5-3)!} = \frac{5!}{2!} = 5 * 4 * 3!$. So if we assume that it's equally likely that we draw any letter, we get a probability of $\frac{1}{5*4*3} = \frac{1}{60}$ for drawing the word "BAR".

- Combinations: $C(n,k) = \binom{n}{k} = \frac{n!}{(n-k)!k!}$. This is similar to the permutations formula, but is a case where *order does not matter*.
- Why might order not matter? Lots of cases where it doesn't. Imagine there are 5 different donuts of different types, and you want to figure out how many different combinations of donuts you could choose, given that you can only eat three (though if they're Dimo's donuts, you should really work harder to increase k...).
- In this case, it doesn't matter if you draw the chocolate one before the sprinkles one. You still have one chocolate donut and one sprinkles donut.
- Example with probability: Imagine donuts are $\{Chocolate, Plain, Sprinkles, Vanilla, Jelly, Powdered, Glazed\}$. If you select two donuts, what is the probability you get the powdered and glazed donuts?
- **Ans:** One combination (Powdered, Glazed) divided by all possible combinations of two ($\frac{7!}{5!2!} = \frac{7*6}{2} = 21$). Gives us $\frac{1}{21}$.
- Bonus question: What's the probability, given that you pick three donuts, that one of them is chocolate?
- **Ans:** This is trickier, but actually pretty intuitive. The denominator is just any combination of three (i.e. $\frac{7!}{4!3!} = \frac{7*6*5}{3*2} = 35$). The numerator is all combinations of three elements that include a chocolate donut. So since order doesn't matter, let's just take the chocolate one out, and think of how many different ways we can choose the two remaining donuts from the remaining elements. So $\frac{6!}{4!2!} = \frac{6*5}{2} = 15$. So our probability of getting a chocolate donut is $\frac{15}{35} = \frac{3}{7}$.
- Side bar on functions and sets: The combinations formula is a two-variable function. As such, it maps elements from one set to elements of another set. What are the domain and co-domain of this function?
- **Ans:** $P(n,k) : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$. Factorials are only defined for non-negative integers, and given two non-negative integer inputs, it produces a single non-negative integer output.

3.5 Introduction to distributions

- In general, random variables map events in the sample space to real numbers.
- In assigning numerical values to objects in the sample space, can simplify the sample space substantially.
- Quick review: What is a sample space?
- We may want to know what the probability is that this variable will take on certain values, or certain intervals of values. This is known as the variable's distribution.
- We can consider this for discrete or continuous random variables.
- For instance, let's consider coin tosses, where the probability of heads is 1/2. Let's define the random variable h as the number of heads. Now let's consider a case where we flip five coins.
- Distribution of this random variable h can be written $P(h = k) = p_h(k) = \binom{5}{k}\ p^k(1-p)^{5-k}$ for $k = 0, 1, \ldots, 5$.
- This is a special case of what's known as the binomial distribution.
- Arbitrary form (note that n and p are parameters, while k is different values that can be assumed by the random variable):
  $P(k; n, p) = \binom{n}{k}\ p^k(1-p)^{n-k}$.
- Notation: We might write this $h \sim Binom(n,p)$

- This is known as the probability distribution function, and for the discrete case, it assigns a probability to each of a finite number of realizations of the random variables. Parameters make it a "family" of distributions.
- In class: find the probability of each realization of the variable for the five coin flip example.
- The cumulative distribution function is the probability that we achieve any value less than or equal to a particular value. E.g. for the case defined above: $F(k) = \sum_{i=0}^{k} \binom{5}{i} p^i (1-p)^{5-i}$ where $p = 1/2$
- To find the probability that the random variable falls between two values, say $2 \leq h \leq 4$ use $F(4) - F(2-1)$. Need to subtract off probability one step below the bottom bound because it's discrete.
- Find cumulative distribution of random variable h.
- Analogous case for continuous random variables, but uses integrals given that the probability of any one value is zero.
- Instead, define probabilities over intervals, e.g. $P([a,b]) = \int_a^b f(t)dt$.
- $f(t)$ is the "density function".
- Given this form, what would the cumulative distribution look like?

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

- Integrals of probability distributions from negative infinity to a number? What does this mean? If there's no support over large parts of this space, then this just adds an area of zero. See next example of uniform distribution.
- Uniform distribution: $f(x) = \frac{1}{b-a}$
- This is a continuous distribution where all points on the interval over which it has support are equally likely.
- Question: Say we have uniform distribution with support of $[-2, 1]$. What is $Pr([-0.5, 0.5])$?
- What is the CDF of the uniform distribution in general terms?
  Ans: $\int_a^x \frac{1}{b-a} dt = \frac{x-a}{b-a}$
- Normal distribution: $f(x; \sigma, \mu) = \frac{1}{2\pi\sigma} e^{-\frac{(x-u)^2}{2\sigma^2}}$
- This is the standard bell curve that you are likely familiar with. It is a continuous distribution function with support across the entire real line.
- σ parameterizes spread of distribution (it's variance) and μ parameterizes position.
- Note: In order for something to be a valid probability distribution or density function, it needs to sum or integrate to 1 when taken over the entire space where it has support. This, essentially, satisfies the axiom that $Pr(S) = 1$.
- Mixed distributions/atoms in the distribution? In formal theory you often see this being ruled out because it's mathematically inconvenient.
- Note: Methods of estimation (e.g. maximum likelihood estimation, or MLE) are about being given data, assuming a process (distribution) that underlies that data, and then trying to determine the parameters of the distribution from that data.
- Closing thought; what does it mean to say that variables are independently and identically distributed?

3.6 More on distributions

- What are some ways that we can connect a linear model to a distribution?
- Say we have a linear model and some cumulative distribution $\phi(x)$ (this could be the standard normal distribution, for instance). We can wrap this distribution around the model by writing it $\phi(X\beta)$
- Models like logit and probit allow us to do this in a way that ensures estimated $\hat{y}$s are between zero and one.

3.7 Expected Values

- In discrete space: $E(x) = \mu = \Sigma_k k \cdot p_x(k)$
- Example of single die, where we want the expected value of rolling a die.
- What about when we have unequal probabilities? Say we flip a coin and add five to the dice total if it lands on heads?
- Practice: $x \sim Binom(4, 0.5)$. What is the expected value of x (i.e. $E(x)$)?
- Analogous continuous case: $E(Y) = \mu = \int_{-\infty}^{\infty} y \cdot f(y)dy$
- This is a measure of central tendency.
- Practice: find expected value of $x \sim U(-1, 3)$.
- Practice: find expected value of $x \sim f(t) = 2t,\ x \in [0, 1]$
- Rules of expectation operator:
  1. $E(a) = a$
  2. $E(bX) = bE(X)$
  3. $E(a + bX) = a + bE(X)$
  4. $\Sigma E(g(X)) = E(\Sigma g(x))$
  5. $E(E(X)) = E(X)$
- Conditional expectation: $E(Y|X)$
- Example: Dice when six has already been rolled. What is conditional expectation of the value?
- Regression function: $E(y|x)$

3.8 Variance and Other Moments

- m^{th} moment of X is $E(X^m)$. m^{th} central moment is $E(X - E(X))^m$
- Variance is second moment.
- Covariance: $E(x - E(x))(y - E(Y))$
- Also can be expressed as $E(X^2) - (E(X))^2$. See proof below.

$$\begin{aligned} Var(X) &= E(X - E(X))^2 \\ &= E(X^2 - 2xE(X) + (E(X))^2) \\ &= E(X^2) - E(2xE(x)) + (E(X))^2 \\ &= E(X^2) - 2E(x)E(x) + (E(X))^2 \\ &= E(X^2) - (E(x))^2 \end{aligned}$$

3.9 Rules of Variance and Covariance

- $Var(a + bX) = b^2 Var(X)$
- $Var(a + bX + cY) = b^2 Var(X) + c^2 VAR(Y) + 2bcCov(X, Y)$
- $Var(c) = 0$
- $Cov(X, Y) = E(XY) - E(X)E(Y)$. Note, this is zero if X and Y are independent, as in this case $E(XY) = E(X)E(Y)$
- $Cov(X + c, Y + b) = Cov(X, Y)$
- $Cov(cX, bY) = cbCov(X, Y)$
- $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$
- $Cov(X, X) = Var(X)$

3.10 Brief Aside: A Superficial introduction to optimization

- What happens when we want to find where a function reaches a maximum or minimum?
- Finding critical points (often maxima and minima). You covered this with Jean?
- Look at function $-x^2 + 4x$. Derivative equals zero when tangent line is horizontal, which in this case corresponds to the maximum.
- What about $x^2 - 4x$?
- How do we determine if maximum or minimum? Intuition versus math. Show instances where it is neither a minimum or maximum.
- If function is concave at critical point (i.e. $f''(x) < 0$) then we have found a maximum. If it is convex $f''(x) > 0$ we have found minimum.
- If $f''(x)$ is undefined or zero, we do not know. Thse are called saddle points.
- Now for an example.

$$\begin{aligned} f(x) &= (x_1 - c)^2 + (x_2 - c)^2 + (x_3 - c)^2 \\ &= x_1^2 + x_2^2 + x_3^2 + 3c^2 - 2cx_1 - 2cx_2 - 2cx_3 \\ \leftrightarrow \frac{df}{dc} &= -2x_1 - 2x_2 - 2x_3 + 6c = 0 \\ \leftrightarrow 3c &= x_1 + x_2 + x_3 \\ \leftrightarrow c &= \frac{x_1 + x_2 + x_3}{3} \end{aligned}$$

- What about optimization when we have a function of many variables?
- Consider $-x^2 + 4x - y^2 + 6x$. Take partial derivatives and set each to zero. What happens if we're below or above the solution values?
- Consider function $x^2 + y^2 - xy - 3y$. Take partial derivatives and set each to zero. Is this a minimum or maximum?
- Important: the optima of a function are stable to monotonic transformations of the function. Any strictly increasing function will do: logarithmic transformations are common, for reasons we will see later.

3.11 Quick Introduction to Maximum Likelihood Estimation

- Up until now, we've been dealing with distributions where we know the value of all the parameters.
- What if we know some things about the distribution (or at minimum, suspect certain things) such as what family of distributions it is part of, but don't know others, such as the value of particular parameters?
- We may, in this case, want to estimate the value of particular parameters.
- Example: Imagine being handed a coin, and not knowing whether or not it's a fair coin. You know that coin flips will be binomially distributed (i.e. you know the family of the distribution) and you know how many times you've flipped it (n) but you don't know p. How might you "guess" p by flipping it a number of times and using the data you generate?
- Any function of the data whose objective is to "guess" the parameter is called an *estimator*, even if this estimator ends up throwing out a lot of the data.
- Computing the value of that function given particular data gives you an *estimate*.
- One approach is known as maximum likelihood estimation (MLE).
- Treat each realization of the coin flip as iid. This allows us to obtain a "likelihood function", which is in effect, the joint probability of your "sample".
- General form: $L(\theta) = \prod_{i=1}^{n} p(k_i; \theta)$, where θ is an arbitrary parameter or set of parameters.
- Intuition: we're just multiplying all the probabilities of the individual observations. For discrete distributions this gives you the joint probability of the sample; for continuous distributions, it gives you the joint density.
- In either case, a sensible approach to estimation is to choose the parameter that maximizes this function, as this is the choice that makes the sample most likely.
- Will often maximize a transformation of the function.
- In particular, recall that logarithms are monotonic functions, so taking the natural logarithm of the likelihood function will not change the optima.
- We can try an example: flip a coin 15 times. How can we use MLE to derive an estimator for p?

Binomial Distribution with $n = 15$
Say we're given $n = 15$, so the only unknown parameter is p. Thus we have:

$$= \binom{15}{x} p^x (1-p)^{15-x}$$

Resulting in the likelihood function (where m is the number of observations)

$$L(p|\boldsymbol{x}, n = 15) = \prod_{i=1}^{m} \binom{15}{x_i} p^{x_i} (1-p)^{15-x_i}$$

And the log-likelihood function (all Σs are $\Sigma_{i=1}^{m}$):

$$logL(p|\boldsymbol{x}, n = 15) = \Sigma x_i log(p) + \Sigma(15 - x_i) log(1-p) + \Sigma log\left(\binom{15}{x_i}\right)$$

Taking the derivative and setting to zero

$$\begin{aligned}
\frac{\partial logL}{\partial p} &= \frac{\Sigma x_i}{p} + \frac{\Sigma 15 - x_i}{1-p}(-1) = 0 \\
\leftrightarrow \frac{1-p}{p} &= \frac{\Sigma(15 - x_i)}{\Sigma x_i} \\
\leftrightarrow \frac{1}{p} - 1 &= \frac{\Sigma(15 - x_i)}{\Sigma x_i} \\
\leftrightarrow \frac{1}{p} &= \frac{\Sigma(15 - x_i) + \Sigma x_i}{\Sigma x_i} \\
\leftrightarrow p &= \frac{\Sigma x_i}{\Sigma(15) - \Sigma x_i + \Sigma x_i} \\
\leftrightarrow p &= \frac{\Sigma x_i}{m(15)}
\end{aligned}$$

- An estimator doesn't need to be good to be an estimator. "Seven" is an estimator. A fair bit of statistics work is on properties of estimators as ways to evaluate them.
- A big divide in political methodology is about how willing people are to make distributional assumptions. Randomized experiments don't require making distributional assumptions for results to hold, but much of the statistical modeling work with observational data does.
- Nonparametric methods try to avoid distributional assumptions, but have less power than parametric (distribution assuming) methods.
- Approaches coming from the causal inference literature (e.g. regression discontinuity designs, natural experiments, etc.) are often about trying to get "as if" random assignment to avoid making distributional assumptions.

4 Linear Algebra

4.1 Vectors

4.1.1 Basics

- *Scalars* are single elements of some set. I.e. is a 1x1 matrix. Other representation: $x \in \mathbb{R}^1$
- *Vectors*, are single dimensional arrays of numbers, i.e. 1xn matrix. Other representation: $x \in \mathbb{R}^n$
- Example $\boldsymbol{x} = \begin{bmatrix} 1 & 3 & 7 \end{bmatrix}$.
- These can also be interpreted as representing points in n-space. Recall that points in R^n are represented by n-tuples, which are vectors. E.g. in R^3 you have 3-tuple $(0, 7, 3)$.
- Can identify elements of a matrix by subscripts. E.g. for above, $x_2 = 3$
- $\mathbf{0} = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix}$
- Matrix extends vector to multiple dimensions.
- Will talk about general matrices later. For now focus on vectors.
- Vector addition just involves adding corresponding elements.
- Scalar multiplication of vectors. If $\boldsymbol{a} = \begin{bmatrix} 1 & 3 & 7 \end{bmatrix}$, what is $c\boldsymbol{a}$?

4.1.2 Dot products, lengths, distances

- **Dot product** of two vectors multiplies corresponding elements and sums each product, i.e. $\boldsymbol{a} \cdot \boldsymbol{b} = \Sigma a_i \cdot b_i$
- For column vectors this is equal to $\boldsymbol{a}'\boldsymbol{b}$
- Who remembers pythagorean theorem? Can determine length of a vector in a similar fashion. Example with 2-vector.
- Distance between two 2-vectors involves subtracting them and then applying Pythagorean theorem.
- *i.e.* between $(1, 2)$ and $(4, 5)$ we would have $\sqrt{(4-1)^2 + (5-2)^2}$
- Helpfully, lengths and distances generalize into n-space in a similar fashion.
- i.e. for R^3 with $(1, 2, 3)$ $(2, 4, 6)$ we would have $\sqrt{(2-1)^2 + (4-2)^2 + (6-3)^2}$
- Take square root of dot product with self for length, i.e. $\|\boldsymbol{a}\| = \sqrt{\boldsymbol{a} \cdot \boldsymbol{a}}$
- For euclidean distance, take $\|\boldsymbol{a} - \boldsymbol{b}\| = \sqrt{(\boldsymbol{a} - \boldsymbol{b}) \cdot (\boldsymbol{a} - \boldsymbol{b})}$

4.1.3 Linear combinations, span of a set

- Linear combinations of vectors: say $c\boldsymbol{a} + d\boldsymbol{b} = \boldsymbol{e}$
- A linearly independent set of vectors is a set where no vector is a linear combination of another.
- **Span of a set of vectors:** the set of all linear combinations of a set of vectors.
- If any vectors in a set are linearly *dependent*, they will not increase the span of the set. Why?
  Ans: Because that vector, and any scalar multiple, could already have been obtained through linear combinations of the other vectors.
- Without getting into a formal treatment of this (598 will do that) one linearly independent vector spans a line (easy to show), two linearly independent vectors will span a plane, three linearly independent vectors span a three dimensional space, etc.
- Indeed, in linear algebra, we will eventually define the "dimension" of a space in terms of the vectors required to span in. Again, this will be addressed in more detail in 598.

4.2 Matrix Algebra

4.2.1 Basics

- Matrices extend vectors to multiple dimensions.
- As with vectors, can multiply a scalar by a matrix. E.g. if $A = \begin{bmatrix} a & b \\ d & e \end{bmatrix}$, then $cA = \begin{bmatrix} ca & cb \\ cd & ce \end{bmatrix}$
- In effect, matrix multiplication involves taking the dot products of rows of one matrix and columns of another, and using those to generate the elements of the new matrix.
- This is why for matrices to be conformable for multiplication you need equal dimensions for columns of one matrix and rows of the other, i.e. $n \times m$ and $m \times j$. This ensures that each *row* of the first matrix is of equal dimensions to each *column* of the second matrix.
- Note that matrix multiplication is not commutative! $\boldsymbol{AB} \neq \boldsymbol{BA}$

- Properties: $(AB)C = A(BC)$
- $(A+B)C = AC+BC$ (keep order of matricies, so that pre and post multiplication works out properly)
- $xAB = (xA)b = A(xB) = ABx$
- Identity matrix (which is matrix with 1s along the diagonal) is such that $\boldsymbol{IA} = \boldsymbol{A}$
- Inverse matrix: $\boldsymbol{AA^{-1}} = \boldsymbol{I}$.
- Question: Can a non-square matrix have an inverse?
- Idempotent matrix: $AA = A$
- Matrix transpose: $\boldsymbol{A}^T$ or $\boldsymbol{A}'$. Switches rows and columns. $\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \leftrightarrow \boldsymbol{A}' = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$
- Important matrix properties: $(A')' = A$
- Additive: $(A + B)' = A' + B'$
- $(AB)' = B'A'$
- $(cA)' = cA'$
- $(A^{-1})' = (A')^{-1}$
- Matrix rank is the number of linearly independent rows or columns of matrix.
- If square matrix rank is not equal to number of rows/columns, then there will be no inverse.

4.2.2 Example of inverse, with R code

```
> mat2
     [,1] [,2] [,3]
[1,]    1    3    3
[2,]    1    4    3
[3,]    1    3    4
> solve(mat2)
     [,1] [,2] [,3]
[1,]    7   -3   -3
[2,]   -1    1    0
[3,]   -1    0    1
```

4.2.3 Determinants, inverses, etc.

- Determinants are a value that can be computed from a square matrix that is useful for finding the inverse.
- Notably, if a square matrix is not full rank, the determinant will be zero.
- In 598 will look at how to compute determinants generally. For now, let's look at 2x2 matrices only.
- For matrix $\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, we have $det(\boldsymbol{A}) = ad - bc$
- Application of a determinant: to find the inverse of a 2x2 matrix, we do:
  $A^{-1} = \frac{1}{det(\boldsymbol{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

- The weird matrix being multiplied by the inverse here is called the "adjoint" matrix. We'll talk about that more in 598, but for now you can just memorize this formula.
- Example: find inverse of $\begin{bmatrix} 1 & 3 \\ 4 & 0 \end{bmatrix}$

4.2.4 Solving Systems of Equations with Matrices (Using Inverses/Cramer's rule)

- Matrices can often be used to represent systems of linear equations, i.e. $Ax = c$
- Example: $\begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x + 2y \\ 3x + 5y \end{bmatrix}$
- We can use inverses to solve systems of equations: $A^{-1}Ax = x = A^{-1}c$
- For the example above, let's choose $\boldsymbol{c} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ and solve using the inverse.
- Can also use something called Cramer's rule, that uses the ratios of determinants of two matrices to find the values for specific variables.
- Example: $x = \dfrac{\begin{vmatrix} 5 & 2 \\ 1 & 5 \end{vmatrix}}{\begin{vmatrix} 1 & 2 \\ 3 & 5 \end{vmatrix}}$
- Replacing column of original coefficient matrix with the values of $\boldsymbol{c}$ for the matrix in the numerator.
- This can be easier when you only care about the value of one variable.
- If you can't compute inverse (e.g. if matrix is not of full rank), also can't solve systems of equations.
- Also can't compute regression estimator (full rank assumption, also known as "no perfect collinearity" assumption)

4.3 ZOMG Regression!!!!11

- Intuition from the basics: what are we doing when we look at a single-variable regression, i.e. a model of form $y = \beta_0 + \beta_1 x_1 + e$?
- Question: So if all we are interested in is the effect of x_1 on y, why don't we just do this all the time?
- Model: Drowning deaths $= \beta_0 + \beta_1$ice-cream sales $+ e$ What's the issue?
- "Lurking" variables, or omitted variable bias. Classic case: where a dependent variable of interest is related to some other dependent variable *and* the independent variable.
- Question: Suppose the true model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. Say we initially estimate $y = \beta_0 + \beta_1 x_1$. Will including x_2 reduce the bias in our estimate of β_1?
- Follow up question: what are the conditions under which the above has different answers?
- Say A causes B causes C. Should we include both A and B?
- In any event, we need a mechanism of "controlling" for omitted variables that we feel may be confounding our analysis.
- Multiple regression does this, in some sense. We "partial out" the effects of other independent variables in order to isolate the effect of a single variable.

- Note: we can obtain estimates for β_1 in a two variable model in a way that illustrates the partialling out interpretation well. Regress x_1 on x_2, obtain the residuals r_1, then regress y on thse residuals r_1. This will give the estimate of β_1 when the effects of x_2 have been partialled out, and is the same as what we would have gotten from doing multiple regression in the first place (though not the same standard errors). See code at end of notes to try it out.
- Predicting bias direction with two variables:

	$Corr(x_1, x_2) > 0$	$Corr(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

- Things get more complicated when the true model includes more than two variables. For instance, say you have x_3 which is uncorrelated with x_1 but is correlated with x_2. Does not including x_3 induce bias in our coefficient for x_1?
- Answer is yes, if x_2 is correlated with x_1. Doesn't matter that x_3 is not directly correlated with x_1.
- Say our correct model is $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Say we start off with just x_1, and say that we know both x_2 and x_3 are correlated with both x_1 and y. Does our bias decrease when we go from including only including x_1 to including x_1 and x_2?
- The answer: not necessarily! This is the subject of Kevin Clarke's wonderfully-titled paper: *The Phantom Menace: Omitted Variable Bias in Econometric Research.*
- If x_2 introduces negative bias and x_3 introduces positive bias, then including only one and not the other means you could be further from the truth than with neither.
- As a result, unless we have the fully specified model, we can't even know if including a variable that belongs in the model with increase or decrease the bias on the coefficient estimate of interest.
- Summary question: You are interested in the effect of x_1 on y. x_2 is also part of the true model. Should you include it?

4.4 OLS estimator derivation

Ordinary least squares linear regression is based on minimizing the squared differences between your regression "line" (hyperplane) and your observed data. Same deal as what we did earlier with least squares estimators for the mean. So, want to minimize $e'e$ where $e = y - XB$ (can you see why this equation holds?)

$$\begin{aligned} min_B\ (y - XB)'(y - XB) &= (y' - B'X')(y - XB) \\ &= y'y - B'X'y - y'XB + B'X'XB \\ &= y'y - 2B'X'y + B'X'XB \end{aligned}$$

taking derivative with respect to B and setting to zero returns

$$\begin{aligned} -2X'y + 2X'XB &= 0 \\ \leftrightarrow X'XB &= X'y \text{ (Note, this is sometimes called the normal equation(s))} \\ \leftrightarrow (X'X)^{-1}X'XB &= (X'X)^{-1}X'y \\ \leftrightarrow B &= (X'X)^{-1}X'y \end{aligned}$$