Political Science Math Camp 2014

Jason S. Davis

August 18-27, 2014

Contents

1 Welcome/Introduction

- Jason Davis - E-mail: jasonsd@umich.edu - Office: Haven Hall 7730 - Telephone: (write in class)
- Math camp objectives: review the building blocks of the content you will be dealing with in 598 & 599 while providing lots of practice along the way.
- All the content will have applications in both statistics and game theory, but the first half (first four days) is directed more towards 598 (i.e. game theory) and the second half (second four days) is directed more towards 599 (i.e. statistics).
- Will cover numbers, functions, basic single-variable calculus, some multivariate calculus in first half; set theory/probability theory and linear algebra in second half.
- In general, a well-pitched class will leave half of people thinking things are too slow and half thinking things are too fast. However, for math camp we will err on the side of going a bit slower.

2 Math in the social sciences

- Why might we want to use mathematics in the social sciences?
- Different approaches: game theory and statistics.
- These are often grouped together as "methods", but there's an extent to which they share little in common besides their use of numbers and symbols.
- Statistics $\rightarrow$ structured empirical investigations. Can be descriptive, or about characterizing "causality".
- Game theory $\rightarrow$ theorizing. Not necessarily empirical; is a collection of logical statements that may be useful or not useful, and may or may not lead us to some expectations about what might occur empirically.

3 Basic building blocks

3.1 Logic

- Will sometimes use logical notation. I.E. $A \rightarrow B$, $A \rightarrow \neg B$, $A \leftrightarrow B$. I don't intend to spend too much time on logic in this course, but it's useful for exposition.
- $A \rightarrow B$. 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.

3.2 Sets and numbers

- Sets are collections of elements. Note: this is different than vectors.
- Different kinds of numbers are sets of all numbers. "Subsets" in particular.
- Examples: $\mathbb{N}$ = natural numbers = $\{1, 2, 3, ...\}$. These include either all positive integers, or all positive integers plus zero.
- $\mathbb{Z}$ = integers = $\{..., -2, -1, 0, 1, 2, ...\}$.
- Can include any kind of elements, and can be characterized through rosters, sets, or intervals.
- Roster: $X = \{A, B, cat, phonebook, 7\}$
- Set-builder: $X = \{X|x = x^2\}$. What is this set?
- Interval notation. $X = (0, 1)$ or $X = (0, 2]$
- Sum, difference, and products of two integers are also integers. Taking quotients allows us to develop set of "rational numbers": $\mathbb{Q}$ = rational numbers = $\{\frac{a}{b}|a, b \in Z, b \neq 0\}$
- Is every number a rational number? No: some like $\sqrt{2}, e, \pi$ are not. These are called "irational numbers".
- Total set of rational and irrational numbers if the set of "real numbers", denoted $\mathbb{R}$.
- Imaginary or complex numbers are also a thing, but we don't tend to deal with them very much
- Another question; what is this set? $X = \{X|x \in \mathbb{R}, x^2 < x\}$.
- Notation: say x is an element of a set Y with $x \in Y$.
- $\forall$ means for all or for each element in something. E.g. $\forall x \in \{1, 2, 3, 4\}, x > 0$.
- $\exists$ means there is at least one element that satisfies a property, e.g. $\exists x \in \{1, 2, 3, 4\} s.t. x < 3$.
- Subsets: $(\forall x \in S, x \in S') \rightarrow S \subseteq S'$
- $(\forall x \in S, x \in S') \wedge (\exists y \in S' s.t.\ y \notin S) \rightarrow S \subset S'$
- Unions combine sets, e.g. if $X = \{A, B, tree, saxophone\}$ and $Y = \{B, saxophone, computer, 9\}$ means $X \cup Y = \{A, B, tree, saxophone, computer, 9\}$.
- Intersections include all elements that are in both sets, i.e. $X \cap Y = \{B, saxophone\}$.
- Draw Venn diagrams.
- How would we write this using the notation before? $(x \in X) \wedge (x \in Y) \rightarrow x \in X \cap Y$
- $A \backslash B$ means elements in A but not B, e.g. $X \backslash Y = \{A, tree\}$.
- What is $\mathbb{Z} \backslash \mathbb{N}$? How about $\mathbb{N} \backslash \mathbb{Z}$? This is null set, or $\varnothing$.
- Other notation: | means such that (I also use "s.t.")
- $\equiv$ means "defined as" or "equivalent to".
- Cartesian products of sets: all ordered pairs of sets.
- Equivalent sets.
- Summation operator: $\Sigma_1^4 x = x + x + x + x$. Can generalize with x_i.
- Product operator: $\Pi_i^4 x = x^4$. Generalize with x_i.

3.3 Exponents

- $x^3 = x \cdot x \cdot x$. E.g. $2^3 = 8$.
- $x^m x^n = x^{m+n}$
- $\frac{x^m}{x^n} = x^{m-n}$.
- $x^{-y} = \frac{1}{x^y}$.
- $x^0 = 1$.
- $(xy)^n = x^n y^n$.
- $(x^n)^m = x^{mn}$
- Practice problem: $((xy)^4)^{0.5} yz^{-3}$
- Square roots are just fractional exponents, e.g. $\sqrt{2} = 2^{1/2}$. Often easier to deal with them in this fashion (in my opinion).

3.4 Introduction to Logarithms and e

- Log is inverse of exponent. Can be written for any base.
- I will focus on logarithms of base e, as these are the most common.
- $e = lim_{n \to \infty}(1 + \frac{1}{n})^n$
- $ln(e) = 1$, $ln(e^2) = 2$, $ln(e^3 23.383) = 323.383$
- $ln(x_1 \cdot x_2) = ln(x_1) + ln(x_2)$. E.g. $ln(e^2 \cdot e^3) = ln(e^3) + ln(e^2) = 5$
- $ln(\frac{x_1}{x_2}) = ln(x_1) - ln(x_2)$
- $ln(x_1^b) = bln(x_1)$. E.g. $ln(e^{12}) = 12ln(e) = 12$
- $ln(1) = 0$
- Logarithms thus can be very useful for dealing with products. E.g. if $y = \Pi_i^n x_i$, $ln(y) = \Sigma_i^n ln(x_i)$.
- First difference of log is also often used to approximate percentage changes. Note that for some variable x_t, $ln(x_t) - ln(x_{t-1}) = ln(\frac{x_t}{x_{t-1}})$

3.5 Inequalities and Absolute Values

- As we've already done a little, we often can represent sets by inequalities.
- Inequalities satisfy many of the same properties we get from equations, but with a few rule changes to consider.
- Keep in mind though that our end outcomes are not single values but sets of values.
- Strict versus weak inequalites. $x > y$ versus $x \geq y$.
- Transitivity: $x > y$ and $y > z$ implies $x > z$.
- Flip signs when multiplying or adding by a negative number.
- Absolute values: distance from zero on number line.

- Absolute value of -6 equals absolute value of 6 for that reason.
- $|x-3|$. Graph it.
- Including both inequalities and absolute values can sometimes be confusing.
- Q: $|x-1|>2$? $|x-1|<2$?
- Ans: $(-\infty,-1)\wedge(3,\infty)$. Second is $(-1,3)$ (interval centered on 1 with 2 on each side)

3.6 Elements of Real Analysis

- Displaying coordinates in 2-space, 3-space, etc.
- R^1, R^2, R^3, R^n. R^1 is the set of real numbers, and n generalizes this to n dimensional space.
- $n=1$ is the number line, $n=2$ is plane, $n=3$ is space.
- Points in R^n are represented by n-tuples, which are vectors. E.g. in R^3 you have 3-tuple $(0,7,3)$
- These are "Euclidean" vectors, i.e. vectors in n-dimensional Euclidean space.
- Can have generalized vectors in general vector spaces, which we may touch on in the linear algebra section, but Euclidean vectors tend to be our domain of interest.
- Addition and subtraction with vectors is simple: subtract or add each element to its corresponding element in other vector. Will deal with other properties in second half of the course.
- Quickly: We might talk about points *within a neighborhood* of another point. This is about describing an interval, disk, or (generalizing to R^n) "ball" around the point.
- If in R^1, we have epsilon interval around c defined as $\{x:|x-c|<\epsilon\}$. Recall that this creates an interval centered on c with ϵ on each side, from inequalities.
- Can generalize easily into 2-space. Anyone remember Pythagorean theorem?
- 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, this generalizes into n-space.
- i.e. for R^3 with $(1,2,3)$ $(2,4,6)$ we would have $\sqrt{(2-1)^2+(4-2)^2+(6-3)^2}$
- An ϵ-ball is the set of points around some point in n-space. i.e. $d(x,y)<\epsilon$ where x and y are vectors in n-space.
- Interior point: point is interior to set X if there exists an ϵ-ball around the point such that all the points in the ball are also in the set.
- Boundary point: Any ϵ-ball includes points both inside and outside the set.
- Open set: Every point in S is an interior point. Still has boundary points, i.e. $x=2$ in $(0,2)$, but these are not included in the set.
- Closed set: Includes all boundary points.
- Bounded: set can be contained in an ϵ-ball.
- Compact (in Euclidean geometry): is both closed and bounded.

3.7 Functions and correspondences

- Functions map elements of some input set to an elementt of another set.
- Simplest functions map R^1 to R^1, are sometimes represented as $f : R \to R$.
- More complicated are two variables functions $f : R^2 \to R$, n-variable functions $f : R^n \to R$, or vector values functions $f : R^n \to R^m$.
- For now, we will start with one-variable functions, move on to multivariable functions, and leave vector-valued functions for some point in the future.
- Domain: input set.
- Co-domain: set the input set is mapped to. **NB: This is often referred to as the range, but this is ambiguous, as range may mean the specific values obtained by the function across the domain**.
- Image: the unambiguous word for which values the function actually obtains over its domain.
- For $f : A \to B$, graph is ordered pairs $G = \{(a, b) \in A \times B | f(a) = b\}$. This is generalization of graphing that we are familiar with.
- Single variable function examples: $f(x) = 2x + 3$. Implies $f(2) = 2(2) + 3 = 7$. We take input of 2 and get output of 7.
- What is the domain, co-domain, and image of this function? How about $f(x) = x^2$? (image is R_+ not R in this second case).
- How about $f(x) = 1/x$? $f(x)\sqrt{x}$?
- Can also define two distinct functions by changing domain and range, e.g. if $f : \mathbb{Z} \to \mathbb{Z}$ is $f(x) = x$.
- Oftentimes infer domain and co-domain from context.
- Linear functions: $y = mx + b$. We love fitting lines to things!
- Polynomial functions, e.g. $y = ax^2 + bx + c$. Degree of polynomial is highest exponent.
- Increasing function: $x > y \to f(x) > f(y)$.
- Decreasing function: opposite.
- Can have strictly increasing and decreasing, or nonstrict.
- Monotonicity means the function is either nondecreasing (i.e. is monotonically increasing) or nonincreasing (i.e. is montonically decreasing).
- Monotone transformations preserve order.
- Local minima and maxima: construct an ϵ-ball such that the value is highest or lowest in that ball.
- Global minima and maxima are if the value is the highest in the image.
- Continuity. Smoothness of function.
- Talk about factoring and graphing functions. Intercepts, etc.
- $g(x)$ is an inverse function of $f(x)$ if and only if $f(g(x)) = x$.
- Inverse functions will sometimes be denoted $f^{-1}(x)$.

- Exists only if function is one-to-one.
- Functions that are one-to-one and "onto" (where onto means that every element of y is paired with an element of x) are called bijections.
- For instance, $\sqrt{x}$ is not onto.
- Composite functions, e.g.: $f(x) = g(x) + h(x)$
- Asymptotes and convergence: e.g. $\frac{1}{1+x}$.
- Maxima and minima at boundaries.

4 Introduction to Differential Calculus (single variable)

- Quick review of limits. Leave formal treatment for 598.
- As before with asymptote, e.g. $\frac{1}{1+x}$. The limit as $x \to \infty$ is defined, but we obviously cannot evalute "at" infinity.
- Not all limits are defined. May increase without bound for example, e.g. $f(x) = x$.
- Slopes: $m = \frac{y_2 - y_1}{x_2 - x_1}$, $y = mx + b$. Standard grade school stuff.
- This helps us determine the rate of change over some interval.
- What about when the function is nonlinear?
- Secant lines give us average rates of change over some interval. Example of distance traveled and speed.
- What if we want to know the rate of change at a particular point (e.g. instantaneous speed)?
- Find the value as the secant interval tends to zero. This is the tangent line. Slightly more formally...
- $\lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$
- Show example with x^2. Class works through x^3.
- Power rule: $(x^n)' = nx^{n-1}$.
- Sum rule: $(f(x) + g(x))' = f'(x) + g'(x)$.
- Production rule: derivative of $f(x)g(x)$ is $f'(x)g(x) + f(x)g'(x)$.
- Constant rule: If a is a constant, $(a)' = 0$
- Derivative of $ln(x)$, $(ln(x))' = \frac{1}{x}$. Derivative of e^x, $(e^x)' = e^x$.
- Chain Rule: Derivative for $f(g(x)) = f'(g(x))g'(x)$.
- Lots of other rules; review on your own, and in PS598.
- Practice: $f(x) = ln(x^{234\pi})$. What is $f'(x)$?
- Different ways of writing derivatives: $f'(x) = f_x(x) = \frac{dy}{dx}$
- Higher order derivatives, e.g. $f''(x)$. Rate of change of rate of change.
- Some functions are not differentiable, or are not differentiable at particular points. These functions are not "well-behaved".

4.1 Using Calculus to Analyze Functions

- Strictly increasing: $\forall x,\ f'(x) > 0$
- Concavity and convexity.
- Concave: any secant line is above f. Formally, for $\lambda \in [0, 1]$, $f(\lambda x' + (1-\lambda)x) \geq \lambda f(x') + (1-\lambda)f(x)$
- Convexity: any secant line between two points on function is below f.
- For $\lambda \in [0, 1]$, $f(\lambda x' + (1-\lambda)x) \leq \lambda f(x') + (1-\lambda)f(x)$
- Strict concavity and convexity changes these inequalities to strict.
- Can use to determine whether function "curves up" or "curves down".
- More generally: a convex combination is a linear combination of points or numbers such that the "coefficients" (in this case, the λs) sum to 1.
- Convex sets contain all convex combinations of points within the set.
- Convex hull is the smallest convex set that contains a set of points.

4.2 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).
- 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 example from the diagnostic.

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

5 Brief introduction to multivariable calculus

- Take derivative with respect to a particular variable in a function. Treat other variables as constants.
- E.g. $f(x, y) = x^2 + y^2 - xy$. $\frac{\partial f}{\partial x} = 2x - y$, $\frac{\partial f}{\partial y} = 2y - x$
- Can also take cross-partial derivatives. This shows how a rate of change changes with another variable.
- Application to interaction effects. Basic partial differentiation is used incorrectly enough that Brambor, Clark, and Golder wrote the most highly cited paper in Political Analysis (2346 citations and counting!) by explaining these.
- In their analysis of APSR, AJPS, and JOP articles from 1998 to 2002, they found that 62% of paper that included interactions did not interpret them correctly, i.e. interpreted coefficients as unconditional marginal effects (which again, is only true with linear model).
- Think of a model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. What is the effect of x_1 on y?
- It's $\frac{\partial y}{\partial x_1}$. In linear model, we have special case where it's equal to β_1.
- 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?

6 Brief introduction to integral calculus

- Indefinite integrals are antiderivatives. Go from $f'(x)$ to $f(x)$
- New power rule: $\int x^n = \frac{x^{n+1}}{n+1}$
- Definite integral $\int_a^b f'(x)dx = f(b) - f(a)$. Often written as $\int_a^b f(x)dx = F(b) - F(a)$
- This is a Riemann integral, it is the limit as the rectangles' widths go to zero (draw on board).
- Formally $lim_{\Delta \to 0} \Sigma_i^N f(x_i)\Delta = \int_a^b f'(x)dx = f(b) - f(a)$
- Useful for lots of things: in statistics, particularly continous distributions.
- Practice problem: Find area under x^2 for $x \in [0, 3]$.
- Practice problem 2: Find area under $x^3 + x$ for $x \in [1, 2]$
- Multiple integrals! Can just evaluate these independently. Also, order of integration is not important (Fubini's Theorem).
- Practice problem: $\int_0^1 \int_0^1 x^2 y \; dxdy$

7 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?

7.1 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 $\bigcap_{i=1}^{n} A_i$.
- Examples with dice.
- Countable versus uncountable sample spaces? To be continued!

7.2 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. $P(B|A) = 0$.
- Bayes Rule used to find conditional probabilities. Will talk about more in a later class (and you'll see this in 598).

7.3 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. Consider that the sample space is now (Prize Location, Door Opened)= $\{(1,3),(2,1),(2,3),(3,1)\}$.
- These respectively have the probabilities $\{1/3, 1/6, 1/6, 1/3\}$. Why?
- Find $P(door2win|door3open)$ and $P(door1win|door3open)$.
- 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.
- Independence: $P(A|B) = P(B)$.

7.4 Quick 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 $k = 0, 1, \ldots, n$.

- This is a special case of what's known as the binomial distribution. Discuss the arbitrary form.
- 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.
- In class: find the probability of each realization of the variable.
- 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: $F(h) = \Sigma_{i=0}^{k} \binom{5}{k} p^k (1-p)^{5-k}$ where $p = 1/2$
- To find the probability that the random variable falls between two values, say $2 \leq h \leq 4$ use $F(4) - F(2-1)$
- 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. $P([a,b]) = \int_a^b f(t)dt$.
- Given this form, what would the cumulative distribution look like?
- Closing thought; what does it mean to say that variables are independently and identically distributed?

7.5 More on distributions

- What does it mean to say that variables are independently and identically distributed?
- 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.
- Integrals of probability distributions from negative infinity to a number? What does this mean? Consider example of uniform distribution: $p(x) = \frac{1}{b-a}$; $\forall x \in [a,b]$, $p(x) = 0$ elsewhere

7.6 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?
- Analogous continuous case: $E(Y) = \mu = \int_{-\infty}^{\infty} y \cdot f(y)dy$
- This, as you know, is the "average", which is a measure of central tendency.
- 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: $E(y|x)$

7.7 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.

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

Rules for Variance

1. $Var(c) = 0$
2. $Var(a + bX) = b^2 Var(X)$
3. $Var(a + bX + cY) = b^2 Var(X) + c^2 Var(Y) + 2bcCov(X, Y)$

8 Matrix Algebra

- *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 $\boldsymbol{x} = \begin{bmatrix} 1 & 3 & 7 \end{bmatrix}$.
- 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.
- Can multiply a scalar by a matrix. E.g. if $A = \begin{bmatrix} a & b \\ d & e \end{bmatrix}$, then $cA = \begin{bmatrix} ca & cb \\ cd & ce \end{bmatrix}$
- 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 $\boldsymbol{a'b}$
- Dot product is zero for vectors that are orthogonal.
- 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)
- $xAB = (xA)b = A(xB) = ABx$
- 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.
- Matrices can often be used to represent systems of linear equations, i.e. $Ax = c$
- Example: $\begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x+2y \\ 3x+5y \end{bmatrix}$
- Identity matrix (which is matrix with 1s along the diagonal) is such that $\boldsymbol{IA} = \boldsymbol{A}$
- Inverse matrix: $\boldsymbol{AA^{-1}} = \boldsymbol{I}$.
- Allows for solving systems of equations using inverses: $A^{-1}Ax = x = A^{-1}c$
- 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 maximum 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.
- If you can't compute inverse, also can't solve systems of equations.
- Also can't compute regression estimator (full rank assumption, also known as "no perfect collinearity" assumption)

8.1 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
```

8.2 Linear regression

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

$$\begin{aligned} -2X'y + 2X'XB &= 0 \\ \leftrightarrow X'XB &= X'y \\ \leftrightarrow (X'X)^{-1}X'XB &= (X'X)^{-1}X'y \\ \leftrightarrow B &= (X'X)^{-1}X'y \end{aligned}$$