Linear algebra notes?

Inner product = multiply each element of two vectors and add each of the results together. If one of the vectors is a stacked (block) vector (say, X), and the other is a vector that consists of scalars (B), then the result will look something like:

B1(X1) + B2(X2) + … Bn(Xn)

The final result will be a vector (say, Y) that corresponds to each unique value of X per variable (e.g., first value of Y will correspond to the first value of X1, X2, … Xn); this is the essential function of regression.

Derivatives:

Functions are commonly represented by f(x), where the value of f is determined by some function of x (x, in this case, is referred to as the “argument). Usually, this value is represented by “y,” with the equation:

y = f(x)

where y is the outcome of some transformation at a (potential) point of x. So if f(x) = x-squared, then y will be whatever value of x is, after it is multiplied by itself (e.g., x = 2, thus y = 4). Easy, right?

In most cases, the function of x (y, or f(x)) can be depicted visually on an x-y graph, where each point of the curve (representing f(x)) corresponds to the value of y, at a specific value of x. f(x) = x-squared, for instance, creates an exponentially increasing curve, starting at 0 (for both sides of the y-axis, since a negative value times itself is always a positive value). In cases where y linearly increases with x (i.e., the curve can be approximated by a straight line), one can easily figure this curve out for themselves by manually inputting values of x into the equation for f(x) (e.g., 1, 2, 3, … 10), draw a “dot” on a graph at the corresponding (x,y) coordinate (again, the value of y is what you get from the equation for f(x)), then draw a straight line through each of those dots. Between each dot, then, should be a specific *rate of change* (e.g., change in y from x = 1 to x = 2), which can be computed easily by *the change in y divided by the change in x* (or, more colloquially; (y2 – y1) / (x2 – x1); this is sometimes referred to as the “rise over run,” which gives the rate at which y increases per unit change in x). So far, so simple.

In cases where f(x) is linear (i.e., you can draw a straight line that intersects perfectly with each of its points), finding the slope of the line is relatively straightforward: pick two (random) points on the graph, take the difference in the values of y (f(x)) and x for each, then divide the change in y (dy, or df) by the change in x (dx). No matter which points you choose, *you will always get the same approximation*. This is the biggest upside to working with linear functions (incidentally, it’s also why much of the world of statistics applies linear functions to relationships between variables, but that’s for another time).

With *non-linear* functions of x, however, this process becomes a lot more complicated. If one were to compute the slope for two different points at one portion of the graph (say, f(x) from x = 5 to x = 6), and then do the same for another portion (say, f(x) from x =2 to x = 3), *the results are very likely to differ from one another*. In fact, for any individual approximation of the slope between two points, the value of the slope (when drawn as a straight line) is likely to differ substantially not only from the function as a whole (i.e., the curve of f(x) across the entire x-y graph), but also the *portion of the function that occurs between those two points.*

Mathematicians have long been aware of this conundrum in attempting to compute the value of the slope for a curve which is explicitly non-linear. After all, the idea itself seems almost paradoxical: if the function f(x) can only be represented by non-linear curve, then trying to compute a slope (which, itself, is commonly conceptualized as a straight line) between any two points on the function seems ludicrous. Nonetheless, math has come a long way over the past few centuries and, as a result, the branch of math known as calculus has provided a means for *approximating* the rate of growth at any individual value of x; specifically, through the utilization of *derivatives*.

Now, to be clear, the notion of a “rate of change at a given point of x” is still (somewhat) paradoxical. By definition, in order for a “change” in y to occur, so too must a change in x. If x does not change, then how, exactly, is one to accept that a rate of change can be discovered, given all we can observe are the values for both x and y at a given point of x? The primary means through which mathematicians get around this issue is by using what are referred to as “limits” (something I still need to read up on), which can be thought of as values (e.g., changes in x) that are so infinitesimally small that they can essentially be thought of as *being equivalent to zero without actually being zero!*

For practical purposes, this has several (pretty insane) applications to approximating the value of a slope for any given point of x under a curve (i.e., the “rate of change at x” for f(x)). Namely, one can calculate the *derivative* of f(x) (sometimes referred to as f’(x), or df / dx \* (x)):

df / dx (x) = (f(x + dx) – f(x)) / dx, OR f’(x) = (y2 – y1) / (x2 – x1)

where f(x + dx) corresponds to the value of f(x) after some infinitesimally small change in x. Here, the math gets a bit algebraic and technical (in terms of actually *solving* this equation for various functions of x, which usually involves the imputation of the function formula in the top half of the equation, and the subsequent elimination of the dx terms), but suffice to say the end result will commonly provide a *concrete formula for finding the slope of a curve at any given value of x*. The derivate of f(x) = x-squared, for instance, comes out to:

d(x-squared) / dx = 2x, OR f’(X) = 2x

meaning, we can substitute any value for x and retrieve the current rate of change at that point. At a value of 3, the rate of change (per unit increase in x) would be 6 (or, y = 6x), while a value of 4 would have a rate of change of 8 (y = 8x). Notice that the value of the derivative (f’(x)) does not remain constant; it changes depending on the value of x. Ergo, the derivative of the function of x (f(x)) is, itself, also a function (specifically, a *function of the function of x*). Seems pretty straightforward, if a bit of a headache to parse out mathematically!

What’s perhaps most interesting about this is that the derivative of f, itself, *can be graphed out*. This is because the derivative behaves identically to any other function, and thus has the exact same properties as a function (e.g., f(x)). One can then plot out the value of the derivative per value of x (with the y axis, in this case, representing f’(x)), which produces a curve (non-linear, of course) that can, if desired, be paired alongside the curve for f(x) (to get a sense as to what degree f’(x) aligns with/diverges from f(x)).

In linear algebra, it appears that some equations utilize the “function of a function” property of f’(x), such that the derivative of f(x) can be thought of as a vector, where each value of f’(x) corresponds to some linear combination of an n-vector (say, a) and the argument of f(x) (i.e., x).

Working out some derivative stuff by hand (specifically, the Taylor Approximation):

f-hat(x) = f(z) + f’(z)(x – z)

where z is a fixed number of some kind (I think). Here, the goal is to discover what f(x) is, with the Taylor approximation giving us of a, well, approximation of f(x) with the above equation (hence the “f-hat” instead of f). What’s interesting about this is that if z and x correspond to same value (say, x = 2 and z = 2), then the Taylor approximation simplifies to:

f-hat(x) = f(z)

Or, in other words, the computed “approximation” of f(x) is equivalent to that which would simply be produced by the underlying function! This, of course, makes sense, given we are able to observe every value of y at each point of x (where f(z) at z = x can be represented by the value of y at x). As z diverges from x, however, the Taylor Approximation strays further away from capturing the true underlying function of x (i.e., true value of y).

As an example, let’s use the derivative for the function y = x-squared (df / dx = 2x). Say z = 4 and x = 3, which can be placed directly into the Taylor Approximation equation:

f-hat(x) = 16 + 8(3 – 4) = 8

Here, f(z) was replaced with 16 since 4^2 is 16, while f’(z) was substituted with 8 since 2 \* 4 is 8. The end result of this computation is an estimated f(x) (again, f-hat), or estimated value of y at z = 4, of 8. Of course, since we know the true underlying function (f(x) = x-squared), we can double-check the result of our Taylor Approximation at z = 4 to what we would actually get at x = 3 using said function:

f(x) = 3^2 = 9

Thus, our Taylor Approximation was off by a value (degree?) of 1 (9 – 8 = 1; note that we would get the same result if z = 2!). Pretty simple, but also what?

Random probability theory stuff:

Crime = a vector A of N size (possible values = 0, 1).

Impulsivity = a vector B of N size (possible values = 0, 1).

P(A) = proportion of sample that has engaged in crime; or total instances of A = 1 / N (say, 20%).

P(B) = proportion of sample that is coded as impulsive; or total instances of B = 1 / N (say, 40%).

P(A and B) = proportion of sample that is both impulsive and has committed a crime (say, 15%).

To find out to what degree the probability of event A is *dependent* on event B (i.e., engaging in crime is dependent on a person being impulsive), we need to first find the odds of being both impulsive and having engaged in crime *when the two events are independent of one another.*

To find this, we compute the (expected) value of P(A and B) under the assumption the two events have no impact on one another, which is:

P(A and B) = P(A) \* P(B) = .20 \* .40 = **.08** (1)

As we can see, the expected probability of being observed as being both impulsive and criminal is .08, which is considerably lower than the observed proportion of impulsive criminals (.15). This suggests that, to some degree, being impulsive is associated with a *higher likelihood of being involved in crime*, relative to what should have been observed if the two were (truly) independent events.

Next, we compute the odds of event A (crime) occurring given B (impulsivity) is observed, assuming that the two events are *not independent* of each other. which is:

P(A|B) = P(A and B) / P(B) = .15 / .40 = **.375** (2)

In doing this, we can see that the probability of being observed as criminal, given one is first observed as being impulsive, is higher than the probability of being observed as criminal for the entire sample. Thus, we seem to have (some degree) of evidence which suggests that the two events are not independent of one another (at least not based on this relatively simplistic test).

To give a counterexample, let’s substitute the expected probability of A and B occurring simultaneously (when both events were assumed to be independent of one another) for the observed probability (i.e., P(A and B)) in the previous equation:

P(A|B) = P(A and B) / P(B) = .08 / .40 = **.20** (notice the switch from .15 to .08)

The result of this is equivalent to the observed proportion of criminals in our sample, meaning our original formula, in instances were events A and B are not dependent on one another, reduces to:

P(A|B) = P(A) (3)

In other words, the probability of being observed as criminal, given a person is first observed as being impulsive, *is equivalent to the probability of being observed as criminal.* This makes sense, as in instances where A and B are independent on one another, being observed as either impulsive or non-impulsive first *should have zero impact on a person’s likelihood of offending.* In this example, however, we computed a (conditional) probability of being observed as criminal, given a person is impulsive, of .375; this implies the two events are (or at least could be) dependent on one another.

Finally, using the formulas provided above, we can also utilize our conditional probability of A given B (i.e., P(A|B)), and our probability of observing B (impulsivity) to find our true observed probability of A and B (i.e., the proportion of impulsive criminals we identified at the start). This is accomplished via a simple rearrangement of equation 2:

P(A and B) = P(A|B) \* P(B) = .375 \* .40 = .**15** (4)

Notice that this equation looks (suspiciously) similar to equation 1 (replacing P(A) with P(A|B)). This is because the former equation (1) assumes independence of events A and B, meaning P(A|B) is reduced to P(A) (as shown in equation 3). When both events are truly independent of each other, equation 1 will (almost) always produce the proportion of overlap between two (dichotomous) variables (or rather, it will correspond to the true likelihood of observing A and B at the same time). When the two events are not independent, however, then equation 1 will produce an inaccurate estimate of the true odds of A and B occurring simultaneously. Instead, we utilize the *conditional probability of observing A (given B)* in place of the probability of observing A, which (should) give us a better understanding of the likelihood of observing A after B has been observed first. This fundamental concept forms the basis of all statistical applications in the social (and physical) sciences.