Derivative Factoids

July 15, 2011

Some people take a full semester (or more) to study derivatives. Almost all of that work is on special cases that we encounter only rarely in applied statistics. Most uses are pretty simple, and you might have trouble reading the books if you don't appreciate some of the simplest things.

1 Definition

Recall the definition of derivative. If $y = f(x)$,

$$\frac{dy}{dx} = \frac{d}{dx}f(x) = lim_{x\to 0}\frac{f(x+\Delta x)-f(x)}{\Delta x}$$

In words, take the values of f at 2 different points, and calculate the slope of a straight line connecting the 2 points. Then make the distance between the points smaller and smaller. If that number exists, it is the derivative.

The function $f(x)$ must be differentiable, which basically means that it is "smooth" enough so you can think about the slope of a tangent line.

2 Linearity

$\frac{d}{dx}a\cdot f(x) = a\cdot\frac{d}{dx}f(x)$

and

$\frac{d}{dx}\{f(x)+g(x)\} = \frac{d}{dx}f(x)+\frac{d}{dx}g(x)$

As a result,

$$\frac{d}{dx}\{a\cdot f(x)+b\cdot g(x)\} = a\frac{d}{dx}f(x)+b\frac{d}{dx}g(x)$$

3 Powers of x.

1. Slope of line.
   If $y = bx$, then
   $\frac{dy}{dx} = b$
   Think backwards for a minute. You usually think of b as a constant, but change gears for a minute to notice

$\frac{dy}{db} = x$
(but that's just a digression)

2. Slope of a square.
   If $y = x^2$, then
   $\frac{dy}{dx} = 2x$
   That is one of the easiest ones to prove and "really believe." Use the definition:

$$\frac{dy}{dx} = lim_{\Delta x \to 0} \frac{(x+\Delta x)^2 - x^2}{\Delta x}$$

$$= lim_{\Delta x \to 0} \frac{x^2 + 2x\Delta x + (\Delta x)^2 - x^2}{\Delta x}$$

$$= lim_{\Delta x \to 0} \frac{2x\Delta x + (\Delta x)^2}{\Delta x}$$

$$= lim_{\Delta x \to 0} \frac{(2x+\Delta x)\cdot \Delta x}{\Delta x}$$

$$= lim_{\Delta x \to 0} 2x + \Delta x$$

$$= 2x$$

   Anyway, I believe in that, and take a lot of the rest on faith (insert smiley face here please).

3. Slope of a cube.
   If y=x^3, then
   $\frac{dy}{dx} = 3x^2$

4. Slope of $y = x^4$.
   $\frac{dy}{dx} = 4x^3$

5. Slope of $y = x^{-1} = 1/x$
   $\frac{dy}{dx} = -1x^{-2}$

6. Slope of $\sqrt{x} = x^{1/2}$
   $\frac{dy}{dx} = \frac{1}{2}x^{-1/2}$

You start to notice a general pattern?

$$\frac{d}{dx}x^N = N \cdot x^{N-1}$$

That is true, whether N is a whole number or a fraction.

4 Logarithms

The only derivative I remember from the top of my head is the one for the natural logarithm.

$$\frac{d}{dx}ln(x) = \frac{1}{x}$$

It can't get any easier than that.

If you have a log to a different base, say $log_{10}()$, the derivative involves a "constant of proportionality." But I never remember it, I always have to look it up in the calculus book.

So wait a minute while I go look in the book.

Ah. Here:

$$\frac{d}{dx}log_b(x) = \frac{1}{ln(b)} \cdot \frac{1}{x}$$

5 Exponentials

Recall Euler's constant, e. It is the base of the natural logarithm. Instead of $ln(x)$ sometimes people write $log_e(x)$, just so you remember the base is the special number. One way to define Euler's constant is by declaring it to be the "magic number such that the derivative of $log_e(x)$ equals 1.

Recall the notation:

$$e^x = exp(x)$$

The derivative is:

$$\frac{d}{dx}e^x = \frac{d}{dx}exp(x) = exp(x)$$

In other words, you "get the same thing back".

As in the case of the logarithm, if you are taking powers of some number besides e then a constant of proportionality enters the picture. The book says

$$\frac{d}{dx}b^x = ln(b) \cdot b^x$$

Note that $ln(e) = 1$, so if you set the base, b, equal to e, then this would reduce to the derivative of $exp(x)$.

6 Derivative of a product

$$\frac{d}{dx}\{g(x) \cdot h(x)\} = \frac{d}{dx}g(x) \cdot h(x) + g(x)\frac{d}{dx} \cdot h(x)$$

or, if you like primes,

$$\frac{d}{dx}\{g(x) \cdot h(x)\} = g'(x)h(x) + g(x) \cdot h'(x)$$

7 Function of a function

The **chain rule** states that:

$$\frac{d}{dx}\{f(g(x))\} = \frac{df}{dx}\mid_{g(x)} \cdot \frac{dg(x)}{dx}$$

That's the derivative of $f(x)$ calculated at the location given by the value $g(x)$, multiplied by the derivative of $g(x)$. Confusing enough? Probably.

Suppose, for example, you had

$$g(x) = x^2$$

and

$$f(x) = ln(x)$$

so

$$f(g(x)) = ln(x^2)$$

$$\frac{d}{dx}f(g(x)) = \frac{1}{x^2} \cdot 2x$$

8 Optimization

You should refer to my other handout entitled "Derivatives and Optimization".

9 What about matrices?

Suppose there is a function that depends on 3 coefficients, as in:

$$S(b_1, b_2, b_3).$$

The first order conditions require that

$$\begin{aligned}\frac{\partial S}{\partial b_1} &= 0\\ \frac{\partial S}{\partial b_2} &= 0\\ \frac{\partial S}{\partial b_3} &= 0\end{aligned}$$

People get bored writing that down, so they use a matrix algebra representation

$$\frac{\partial S}{\partial b} = 0$$

The thing on the left is a 3 element column vector, the thing on the right is a column vector with 3 0's.

In all honesty, my mind does not "do calculus" in matrices. If I have to do any calculations, I always end up writing in out, and then converting back to matrices. Maybe if I got more practice (and training), I wouldn't be so limited.

In any case, there are some “derivative factoids” that apply if you work in matrix algebra. These “factoids” are really just adaptations of the things we know are true in calculus.

If you look at Myers, Montgomery, and Vining, you see that the matrix algebra is not really the “work horse.” Rather, it is the “show horse.” If they want to explain something that you can understand, they write out several equations. If they want something concise, they use matrix algebra.

For example, consider p. 73. The First Order Conditions for the maximum likelihood problem, the so-called “score equations”, will have 1 equation for each parameter. Ssuppose there are 3 parameters, $\beta = (\beta_1, \beta_2, \beta_3)$. The first order conditions are given by 3 equations, one for each parameter to be estimated. Each one is a sum over N observations, and each term in the sum is the “prediction error” for the observation, $[y_i - f(x_i,\beta)]$ multiplied by the derivative of the predictive equation for that case, $\frac{\partial f(x_i,\beta)}{\partial \beta_j}$

$$\begin{array}{rcl} \frac{1}{\sigma^2}\sum_{i=1}^N [y_i - f(x_i,\beta] \frac{\partial f(x_i,\beta)}{\partial \beta_1} & = & 0 \\ \frac{1}{\sigma^2}\sum_{i=1}^N [y_i - f(x_i,\beta] \frac{\partial f(x_i,\beta)}{\partial \beta_2} & = & 0 \\ \frac{1}{\sigma^2}\sum_{i=1}^N [y_i - f(x_i,\beta] \frac{\partial f(x_i,\beta)}{\partial \beta_3} & = & 0 \end{array} \quad (1)$$

It gets pretty boring typing that over and over. So they put it into matrix form to save some energy. As usual, let

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \\ \vdots \\ \\ \\ y_N \end{bmatrix}$$

Further, let the column of predicted values be

$$\hat{\mu} = \begin{bmatrix} \widehat{\mu_1} \\ \widehat{\mu_2} \\ \\ \vdots \\ \\ \\ \widehat{\mu_N} \end{bmatrix} = \begin{bmatrix} f(x_1,\beta) \\ f(x_2,\beta) \\ \\ \vdots \\ \\ \\ f(x_N,\beta) \end{bmatrix}$$

The slope of the predicted value for each case for each variable is given by a matrix that has individual elements like this.

$$D_{ij} = \frac{\partial f(x_i,\beta)}{\partial \beta_j}$$

That’s for the i’th case and the j’th variable. If you collected all of those, you’d have an

Nx3 matrix, as in

$$D = \begin{bmatrix} D_{11} & D_{12} & D_{13} \\ D_{21} & D_{22} & D_{23} \\ \vdots & & \vdots \\ \\ D_{N1} & D_{N2} & D_{N3} \end{bmatrix} = \begin{bmatrix} \frac{\partial f(x_1,\beta)}{\partial \beta_1} & \frac{\partial f(x_1,\beta)}{\partial \beta_2} & \frac{\partial f(x_1,\beta)}{\partial \beta_3} \\ \frac{\partial f(x_2,\beta)}{\partial \beta_1} & \frac{\partial f(x_2,\beta)}{\partial \beta_2} & \frac{\partial f(x_2,\beta)}{\partial \beta_3} \\ \vdots & & \vdots \\ \\ \frac{\partial f(x_N,\beta)}{\partial \beta_1} & \frac{\partial f(x_N,\beta)}{\partial \beta_1} & \frac{\partial f(x_N,\beta)}{\partial \beta_1} \end{bmatrix}$$

And so the 3 equations specified above in 1 can be written with matrices as

$$\frac{1}{\sigma^2} D'(y - \hat{\mu})$$

That is considerably more concise. Note that if you had a linear model, one for which

$$f(x_i,\beta) = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Then $\frac{\partial f}{\partial \beta_j} = x_{ij}$, and so the matrix D is just the same old “data matrix” you are used to.

$$D = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots \\ \\ x_{N1} & x_{N2} & x_{N3} \end{bmatrix}$$