POLS509 - The Linear Model
January 23, 2003
More Calculus

1 More Differential Calculus

1.1 Higher-Order Derivatives

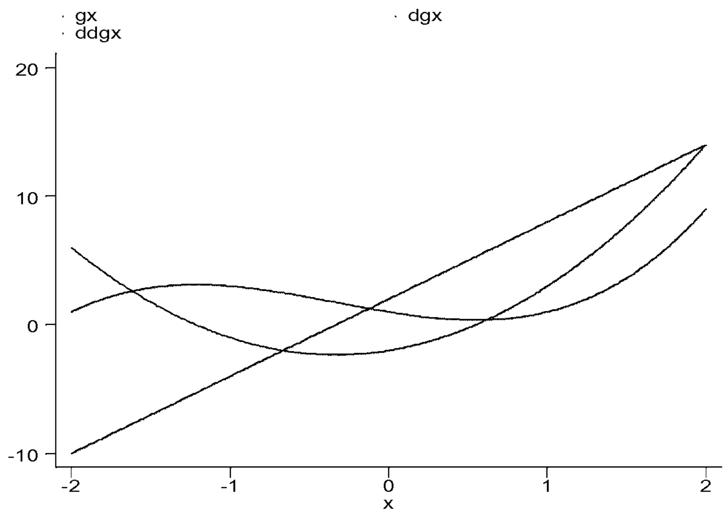
Second derivatives are what we get when we take the derivative of a derivative; they are written as $\frac{\partial^2}{\partial X^2} f(X)$.

For example, consider the function $\mathbf{g(X) = X^3 + X^2 - 2X + 1}$.

- The first two derivatives are:
    - $\frac{\partial}{\partial X} g(X) = 3X^2 + 2X - 2$.
    - $\frac{\partial^2}{\partial X^2} g(X) = 6X + 2$.
- Calculate a few points:

X	$g(X)$	$\frac{\partial}{\partial X} g(X)$	$\frac{\partial^2}{\partial X^2} g(X)$
-2	-1	6	-10
-1	1	-1	-4
0	-1	-2	2
1	-1	3	8
2	7	14	14

- Graph them...



Here,

- $\frac{\partial}{\partial X} g(X)$ is the equation defining slope of the line tangent to $g(X)$ at X. Its a curve, since the slope changes...
- $\frac{\partial^2}{\partial X^2} g(X)$ is the equation defining the slope of the line tangent to $\frac{\partial}{\partial X} g(X)$ at X. Its a line, since the slope of $\frac{\partial}{\partial X} g(X)$ is changing at a constant rate.

1.2 Why should we care?

(About derivatives, that is...).

One answer is, we can use derivatives to find **maxima** and **minima** of functions...

The maximum of some function is the value of X at which $f(X)$ takes on its largest value, while the minimum is where $f(X)$ is lowest.

Q: What ought to happen to the derivative of a function at a minimum or a maximum?

A: It will be zero, because the slope of the line tangent to the function at that point will be horizontal (or, alternatively, its instantaneous rate of change will be zero at that point).

Finding minima/maxima involves *three steps*:

1. Take the first derivative of a function.
2. Set that derivative equal to zero.
3. Solve the resulting equation.

1.3 Example: My love for coffee...

If X is the number of cups of coffee I consume in 24 hours, my utility for coffee is:

$$U(X) = -X^2 + 16X + 10 \tag{1}$$

Q: What is the optimum number of cups of coffee for me to drink in a day?

A: First, take the derivative of the utility function:

$$\frac{\partial}{\partial X} U(X) = -2X + 16$$

Then set it to zero, and solve for X:

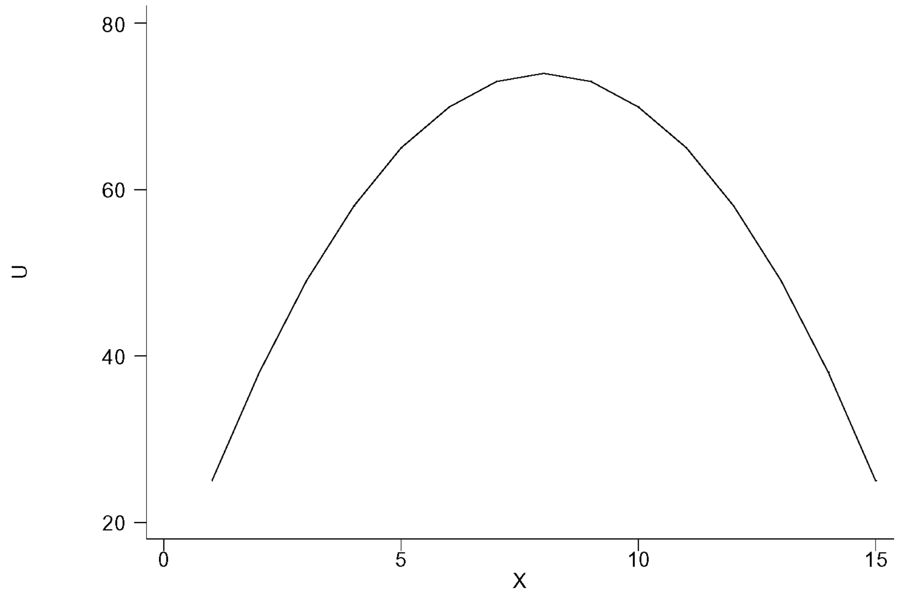
$$\begin{aligned} -2X + 16 &= 0 \\ -2X &= -16 \\ X &= 8 \end{aligned}$$

So $X = 8$ is a *critical point*; i.e. a point at which the point reaches a minimum or maximum.

How do we know if its a minimum or a maximum?

(At least) Three possibilities:

1. We can graph the function itself, and look at it...



2. We can plug values near $X = 8$ into $U(X)$, and see if they yield values less than or greater than that obtained by $X = 8$:

X	$U(X)$
7	73
8	74
9	73

3. We can take the *second derivative*, and evaluate it at the critical point.

   - If $\frac{\partial^2}{\partial X^2}U(X)$ is negative, then the critical point is a maximum.
   - If $\frac{\partial^2}{\partial X^2}U(X)$ is positive, then the critical point is a minimum.

Why?

- Recall that $\frac{\partial^2}{\partial X^2}f(X)$ is the function which maps the slope of the line tangent to the derivative of the function to X.
- If the slope of the second derivative is positive, it means that the series of slope–of–tangents which compose the first derivative are increasing in value in X. That is,
  - As X increases, the slope of the lines tangent to the function at X grows less negative (i.e., starts out sloping down, winds up sloping up).

- E.g. the function $g(X) = X^3 + X^2 - 2X + 1$, given earlier.
- This is indicative of the contour of the function at that point.

- The reverse is true of functions which are at their maximum...

Here, $\frac{\partial^2}{\partial X^2}U(X) = -2$, so the point $X = 8$ is a maximum.

1.4 Multivariate Functions and Partial Derivatives

So far, we've only talked about *univariate* functions – functions that contain only one variable. A *multivariate* function is simply a function which maps from two or more variables to $\mathbb{R}$.

- We write (e.g.) $f(X, Y) = ...$
- A k–variate function can be plotted in $k + 1$–space.
- You already know something about multivariate functions, whether you realize it or not...(e.g., multiple regression models).

Partial derivatives are used to determine the minima and maxima of multivariate functions...

- We write (e.g.) $\frac{\partial}{\partial X}f(X, Y)$ and $\frac{\partial}{\partial Y}f(X, Y)$.
- We say, "The partial derivative of $f(X, Y)$ with respect to X (or Y)."

Calculating partial derivatives is easy: simply treat any variables other than those you are interested in as constants (so that they "drop out" when you take the derivative w.r.t. the variable of interest).

1.4.1 Higher–Order Partials

One can also take second (and higher) partial derivatives, by simply taking additional derivatives of the first-order partial. These are simply partials of partials, and are written as $\frac{\partial^2}{\partial X^2}f(X, Y)$, $\frac{\partial^2}{\partial Y^2}f(X, Y)$, etc.

Finally, we can take what are called mixed partials, or cross–partials. These involve taking the first–order partial of a function with respect to one variable, and then taking the second–order partial of that with respect to the second variable. Here, notation is important:

- $\frac{\partial^2}{\partial X \partial Y}f(X, Y)$ means "First take the partial of $f(X, Y)$ with respect to X, and then take the partial derivative of that with respect to Y."
- $\frac{\partial^2}{\partial Y \partial X}f(X, Y)$ means the reverse: "First take the partial of $f(X, Y)$ with respect to Y, and then take the partial derivative of that with respect to X."

1.4.2 Multivariate Maximization

Cross–partials allow us to determine maxima and minima for multivariate functions, in a similar three–step process to that for univariate functions:

1. Take the first–order partials with respect to each of the k variables of interest, set them to zero, and solve the system of k equations to determine the function's "critical points" in those variables. These points are the necessary conditions for an extremum (i.e. a "flat" tangent). So, for a bivariate function $f(X, Y)$, call these critical points (a, b).

2. Go on to calculate the second–order partials, and the cross–partials, for each of the variables.

3. From these, calculate the *discriminant* at the critical values. For the bivariate function mentioned a bit ago, the discriminant at a and b is:

$$D(a,b) = \left[\frac{\partial^2}{\partial X^2_{(a,b)}} \times \frac{\partial^2}{\partial Y^2_{(a,b)}}\right] - \left[\frac{\partial^2}{\partial X_{(a,b)} \partial Y_{(a,b)}}\right]^2 \tag{2}$$

   We can use the discriminant to determine whether a minimum or maximum exists at (a, b), and whether it is a minimum or maximum, in a fashion analogous to the "second derivative test" for a single variable:

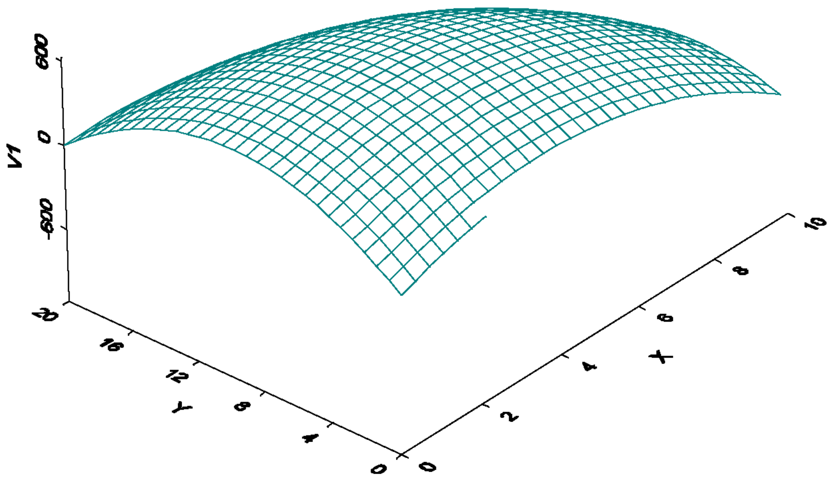
   - If $D(a, b) < 0$, then (a, b) is not a critical point.
   - If $D(a, b) = 0$, then the test gives us no information.
   - If $D(a, b) > 0$, then we have to check out look at the second-order partials...
     - If $\frac{\partial^2}{\partial X^2} > 0$, then the point is a minimum.
     - If $\frac{\partial^2}{\partial X^2} < 0$, then the point is a maximum.

1.4.3 Another Example: Coffee and Cigarettes

Back in graduate school, I used to smoke. In a given 24–hour period, my utility for coffee (X) and cigarettes (Y) was:

$$U(X, Y) = -16X^2 - 5Y^2 + 128X + 100Y + 4XY \tag{3}$$

FYI, that function looks like this:



Given this, **what is the optimal number of cups of coffee and cigarettes I should have consumed in a day?**

To answer this question, we have to maximize this bivariate function.

Step One: Calculate the two partial derivatives.

$$\frac{\partial}{\partial X}U(X,Y) = -32X - 4Y + 128 \tag{4}$$

$$\frac{\partial}{\partial Y}U(X,Y) = -10Y - 4X + 100 \tag{5}$$

Step Two: Set these two equations to zero and solve.

$$\begin{aligned} -32X - 4Y + 128 = 0 \quad &\Leftrightarrow \quad -32X = 4Y - 128 \\ &\Leftrightarrow \quad X = -\frac{1}{8}Y + 4 \end{aligned}$$

$$
\begin{aligned}
-10Y - 4X + 100 = 0 \quad &\Leftrightarrow \quad -10Y - 4[-\frac{1}{8}Y + 4] + 100 = 0 \\
&\Leftrightarrow \quad -10Y + \frac{1}{2}Y - 16 + 100 = 0 \\
&\Leftrightarrow \quad -10.5Y = -84 \\
&\Leftrightarrow \quad \mathbf{Y = 8}
\end{aligned}
$$

$$
\begin{aligned}
-32X - 4Y + 128 = 0 \quad &\Leftrightarrow \quad -32X - 4(8) + 128 = 0 \\
&\Leftrightarrow \quad -32X = -96 \\
&\Leftrightarrow \quad \mathbf{X = 3}
\end{aligned}
$$

So, we know (or, at least, suspect) that the point $(3, 8)$ is a critical point.

Step Three: Calculate the second derivatives, and the cross–partials.

First, the second derivatives:

$$\frac{\partial^2}{\partial X^2} U(X, Y) = -32$$

$$\frac{\partial^2}{\partial Y^2} U(X, Y) = -10$$

Next, the cross–partials (turns out they are the same...):

$$\frac{\partial^2}{\partial X \partial Y} U(X, Y) = -4$$

$$\frac{\partial^2}{\partial X \partial Y} U(X, Y) = -4$$

Step Four: Evaluate the discriminant.

This is also not too hard, since none of the higher–order things vary...

$$
\begin{aligned}
D(3, 8) &= \left[\frac{\partial^2}{\partial X^2_{(a,b)}} \times \frac{\partial^2}{\partial Y^2_{(a,b)}} \right] - \left[\frac{\partial^2}{\partial X_{(a,b)} \partial Y_{(a,b)}} \right]^2 \\
&= [(-32) \times (-10)] - (-4)^2 \\
&= 320 - 16 \\
&= \mathbf{304}
\end{aligned}
$$

What have we learned?

- $D(3,8)$ is positive, which means that $(3,8)$ is a critical point,
- Both second–order partials are negative, meaning that $(3,8)$ is a *maximum* of the function.
- Eight cigarettes and three cups of coffee a day were bliss for me in grad school...[1]

[1] Also, not surprisingly, cigarettes and coffee are substitutes. For example, if I didn't drink any coffee, my utility was maximized by smoking 11 cigarettes; on the other hand, if I drank nine cups of coffee, my optimal cigarette consumption was only six...

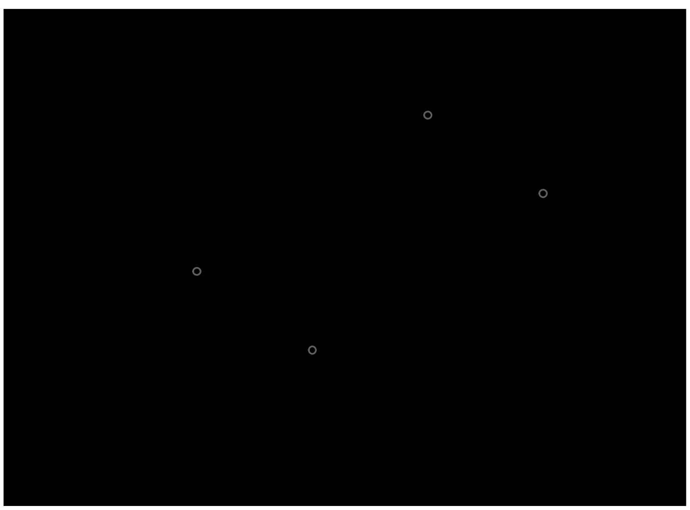
1 Example: A Two-Variable Optimization Problem[1]

Consider a function: $y_i = f(a, b) = a + bx_i$. (Doesn't this look familiar?)

You have a small dataset of observations $i \in \{1, 2, 3, 4\}$.

i	x_i	y_i
1	1	2
2	2	1
3	3	4
4	4	3

The scatterplot of y_i against x_i looks like this:



.

Your task: find the values of a and b that minimize the sum of squared residuals from the line, where a "residual" is

$$e_i = y_i - f(a, b) = y_i - (a + bx_i) = y_i - a - bx_i.$$

[1](Comparable to Hagle's pps. 54-56.)

The function we want to minimize is thus

$$
\begin{aligned}
& \sum_{i=1}^{4} (y_i - a - bx_i)^2 \\
= \;& (2-a-b)^2 + (1-a-2b)^2 + (4-a-3b)^2 + (3-a-4b)^2 \\
= \;& (4 - 4a + a^2 - 4b + 2ab + b^2) + ... \\
= \;& 30 + 4a^2 - 20a + 20ab - 56b + 30b^2.
\end{aligned}
$$

To minimize this, we take the first partial derivatives like so:

$$
\begin{aligned}
\frac{\partial f(a,b)}{\partial a} &= 8a - 20 + 20b \\
\frac{\partial f(a,b)}{\partial b} &= 20a - 56 + 60b,
\end{aligned}
$$

we set them equal to zero and solve these resulting first-order conditions:

$$
\begin{aligned}
FOC(1) &: \; 8a - 20 + 20b = 0 \\
FOC(2) &: \; 20a - 56 + 60b = 0.
\end{aligned}
$$

Starting with FOC(1),

$$
\begin{aligned}
8a &= 20 - 20b \\
8a &= 20(1-b) \\
a &= \frac{20}{8}(1-b).
\end{aligned}
$$

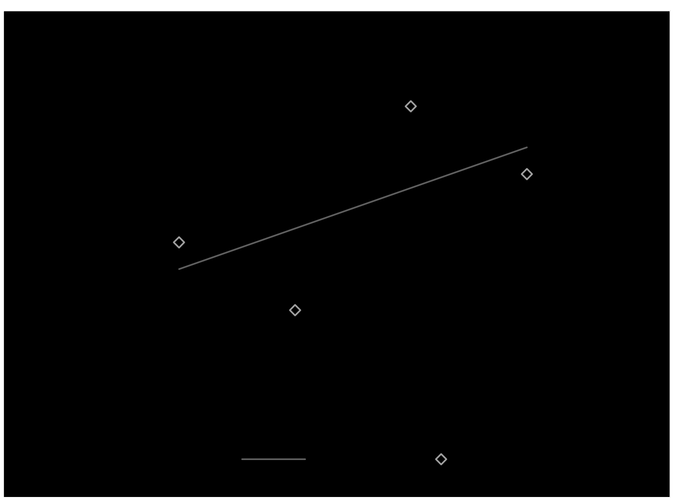
Then substitute into FOC(2) and solve for b:

$$
\begin{aligned}
20a - 56 + 60b &= 0 \\
20(\frac{20}{8}(1-b)) - 56 + 60b &= 0 \\
-50b + 60b &= -50 + 56 \\
10b &= 6 \\
b &= \frac{6}{10}.
\end{aligned}
$$

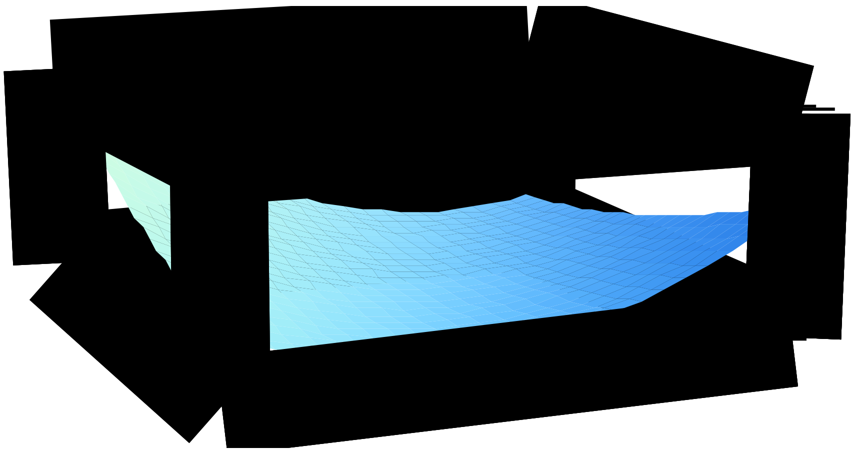
Then plug $b = \frac{6}{10}$ back into either FOC and solve for a:

$$
\begin{aligned}
8a - 20 + 20b &= 0 \\
8a - 20 + 20(\frac{6}{10}) &= 0 \\
8a &= 20 - 12 \\
a &= 1.
\end{aligned}
$$

Excellent! We have found the critical points for this minimization problem: $(a = 1, b = 0.6)$. The function $f(a, b)$ then yields the following line plot:



However, we have to finish the remaining step in the problem: check that these critical points do indeed yield a minimum rather than a maximum. We can check this graphically, plotting the objective function (the sum of squared residuals) like so:



(b is on the axis going right to left)

This indeed looks like a minimum. But let's check this using the discriminant. Our second-order conditions are:

$$
\begin{aligned}
\frac{\partial^2 f(a,b)}{\partial a^2} &= 8 \\
\frac{\partial^2 f(a,b)}{\partial b^2} &= 60 \\
\frac{\partial^2 f(a,b)}{\partial a \partial b} &= 20 \\
\frac{\partial^2 f(a,b)}{\partial b \partial a} &= 20
\end{aligned}
$$

(Note the two cross-partials will always be equal.) The discriminant $D(a,b)$ is thus

$$
\begin{aligned}
D(a,b) &= [8 \times 60] - [20 \times 20] \\
&= 480 - 400 \\
&= 80.
\end{aligned}
$$

Because $D(a,b) > 0$ and $\frac{\partial^2 f(a,b)}{\partial a^2} > 0$, then we have a local minimum at (a,b). Cool!

POLS509 - The Linear Model
January 28, 2003
Even More Calculus

1 Integral Calculus

1.1 The Intuition

Calculus is often broken down into differential and integral calculus, and the two are closely related. You might (as Hagle does) think of integration as the "opposite" of differentiation (a la addition/subtraction, etc.). In many respects, however, its easier to think of integral calculus in terms of **area**...

Q: How do we calculate the area of an irregular polygon? (A: subdivide it.)

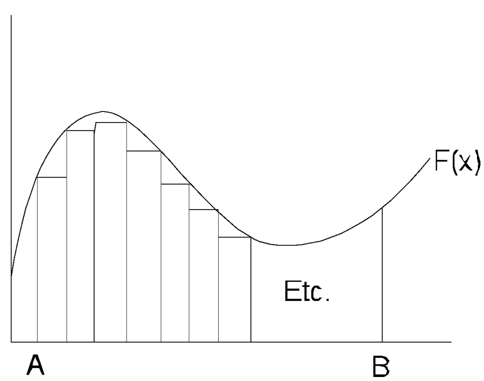
Q: What about the area under a *curve*?...

Consider the region of a plane defined by the X-axis, vertical lines at a and b, and some function $f(X)$.

- Assume that $f(X)$ is continuous and differentiable over this range.
- We want to know the area **A** under the graph of f from a to b. How do we do it?

Following our geometry example, above, we could subdivide it into rectangles...

- One rectangle will be a bad approximation of the area,
- Two rectangles will be better, but still bad,
- Four rectangles will be better still, etc.



More generally:

1. Divide up the interval $[a, b]$ into N subintervals of the same length $\frac{(b-a)}{N}$.
   - Call this value $\Delta X = \frac{(b-a)}{N}$.
   - Choose numbers $X_0 ... X_N$ as follows:
     - $X_0 = a$,
     - $X_N = b$,
     - $X_i - X_{i-1} = \frac{(b-a)}{N}, i \in \{1, 2, 3, ...N\}$.
2. Since the function is continuous, for each interval we know that the function achieves some minimum value $f(u_i)$ at some particular point u_i in the interval $[X_{i-1}, X_i]$.
3. We can then draw a rectangle of width $\frac{(b-a)}{N}$ and height equal to $f(u_i)$ for every X_i...
4. The area of each of these rectangles is simply their width (ΔX) times their height ($f(u_i)$), that is, $A_i = f(u_i)\Delta X$.
5. The total area under the curve, then, is just the sum of these areas:

$$\mathbf{A} = \sum_{i=1}^{N} A_i = \sum_{i=1}^{N} f(u_i)\Delta X$$

Now, to get a better and better approximation of the area $\mathbf{A}$, we can simply make ΔX smaller and smaller, so long as ΔX doesn't equal zero... That is, we might think of the area under the curve $f(X)$ between a and b as:

$$\mathbf{A} = \lim_{\Delta X \to 0} \sum_{i=1}^{N} f(u_i)\Delta X$$

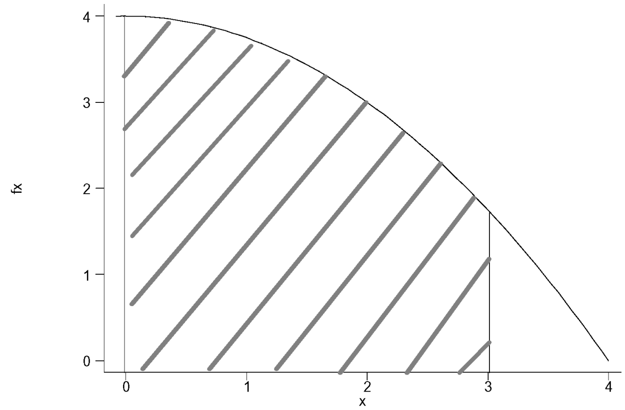
This is, more or less, the definition of a (definite) integral – more on that in a minute...

1.2 An Example

Let's try this out...

If $f(X) = 4 - 0.25X^2$, calculate the area of the region under the graph of $f(X)$ from 0 to 3.

1. First graph the function...



2. If we divide the region into N subintervals, each will have a width of $\frac{(3-0)}{N} = \frac{3}{N}$.

3. $f(X)$ is monotonically decreasing on the range $[0, 3]$; this means that the minimum for each of the N subintervals we create will be at the largest value of X in the subinterval. In other words, $u_i = \frac{3i}{N}$.

4. This in turn means that the height of the function at u_i is:

$$\begin{aligned} f(u_i) &= f\left(\frac{3i}{N}\right) \\ &= 4 - 0.25\left(\frac{3i}{N}\right)^2 \\ &= 4 - 0.25\left(\frac{9i^2}{N^2}\right) \\ &= 4 - \left(\frac{9i^2}{4N^2}\right) \end{aligned}$$

5. The summation of these terms is then equal to:

$$\begin{aligned}
A = \sum_{i=1}^{N} f(u_i)\Delta X &= \sum_{i=1}^{N} \left[4 - \left(\frac{9i^2}{4N^2}\right)\left(\frac{3}{N}\right)\right] \\
&= \sum_{i=1}^{N} \left(\frac{12}{N} - \frac{27i^2}{4N^3}\right) \\
&= \sum_{i=1}^{N} \frac{12}{N} - \sum_{i=1}^{N} \frac{27i^2}{4N^3} \\
&= N\frac{12}{N} - \frac{27}{4N^3}\sum_{i=1}^{N} i^2 \\
&= 12 - \left(\frac{27}{4N^3}\right)\left[\frac{N(N+1)(2N+1)}{6}\right] \\
&= 12 - \left(\frac{9}{8N^3}\right)(2N^3 + 3N^2 + N)
\end{aligned}$$

6. If we recall that $\Delta X = \frac{(b-a)}{N}$, its clear that the way to make $\Delta X \to 0$ is to let N increase without bound. So:

$$\begin{aligned}
\lim_{N\to\infty} \sum_{i=1}^{N} f(u_i)\frac{(b-a)}{N} &= \lim_{N\to\infty} \left[12 - \left(\frac{9}{8N^3}\right)(2N^3 + 3N^2 + N)\right] \\
&= 12 - \frac{9}{8}\lim_{N\to\infty} \frac{2N^3 + 3N^2 + N}{N^3} \\
&= 12 - \frac{9}{8}\lim_{N\to\infty} \left(2 + \frac{3}{N} + \frac{1}{N^2}\right) \\
&= 12 - \frac{9}{8}(2 + 0 + 0) \\
&= 12 - \frac{9}{4} \\
&= \frac{39}{4} \approx 9.75
\end{aligned}$$

(Whew!)

This is the intuition behind integral calculus.

1.3 A Definition

Based on all this, we can say that:

> **If a function f is defined on a closed interval $[\mathbf{a}, \mathbf{b}]$ in $\mathbf{X} \in \mathbb{R}$, the definite integral of f from a to b is given by**:

$$\int_a^b f(X)\,dX = \lim_{P\to\infty} \sum_{i=1}^{N} f(u_i)\Delta X_i$$

Note a couple changes in the notation here...

- They symbol $\int$ is an *integral sign*.
- a and b are referred to as the *limits of integration*.
  - They refer to the range over which the function is integrated.
  - a is the lower limit, b is the upper limit.
- $f(X)$ is called the *integrand*.
- dX indicates the *variable*.
- P represents the largest value of the partition of the function into "pieces": as before, we partition the function into $i = 1, 2, ...N$ "pieces"; letting P go to zero implies that the partitions get smaller and smaller.
- u_i is (again) the minimum value of the function in a particular partition i

1.4 Properties Of Definite Integrals

- If a function is continuous on $[a, b]$ then it is integrable on $[a, b]$.
- $\int_a^b f(X)\,dX = -\int_b^a f(X)\,dX$ – that is, interchanging the limits of integration changes the sign of the integral.
- $\int_a^a f(X)\,dX = 0$ – integrating from a to a equals zero. (Intuitively, there is no "area" to a vertical line....).

1.5 Indefinite Integrals

These deal with the connection between integrals and derivatives. Hagle begins by immediately talking about integrals as the "opposite" of derivatives (which are sometimes called "antiderivatives"). Integration which is the "opposite" of differentiation yields **indefinite integrals**.

Generally, a function $F(X)$ is an indefinite integral of another function $f(X)$ if

$$\frac{\partial}{\partial X}F(X) = f(X)$$

1.5.1 For Example...

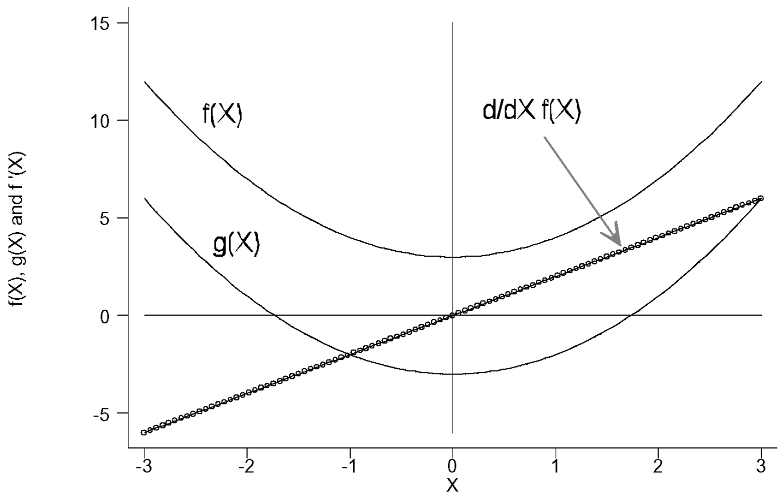
If we take the function $F(X) = X^2 + 4X - 8$, its derivative is $f(X) = 2X + 4$. Conversely, the integral of $f(X) = 2X + 4$ is $X^2 + 4X + c$.

Note that, because the constant terms "drop out" when we differentiate, we can't "recover" them when we reverse the process and integrate.

- Put differently, there are an infinite number of integrals for the function $2X + 4$ (the integral is not *unique*).
- This is because the derivative of a function tells us nothing about the "location" of the function in the space vis–a–vis the Y–axis.
- Hence, "indefinite"...

Consider two functions:

$$f(X) = X^2 + 3 \qquad\qquad g(X) = X^2 - 3$$



- Q: What's the difference in their graphs? A: The location of the graph on the Y–axis...
- The slope of the lines tangent to these functions are the same; therefore they have the same indefinite integral...
- In other words, any function $X^2 + c$ would have the same derivative ($2X$).
- Note that this also implies the reverse: that the integral of $2X$ can be any function of the form $X^2 + c$.

1.5.2 Rules For Indefinite Integrals

$$\int X^n \, dX = \frac{X^{n+1}}{n+1} + c \qquad (1)$$

- This is the power rule for integration.
- Note that

$$\frac{\partial}{\partial X}\left(\frac{X^{n+1}}{n+1} + c\right) = X^n$$

- Also, by convention, $\int X^{-1} \, dX = ln|X| + c$

$$\int dX = X + c \qquad (2)$$

- The integral of a constant is $X + c$.

$$\int [a \times f(X)]\, dX = a \int f(X)\, dX \tag{3}$$

- You can "pull a constant through" an integral.

$$\int [f(X) + g(X)]\, dX = \int f(X)\, dX + \int g(X)\, dX \tag{4}$$

- *Addition*: The integral of sums equals the sums of the integrals.

Indefinite integrals will come back to haunt us when we begin talking about probabilities and probability distributions a bit later...

2 Probability And Probability Theory...

2.1 Random Variables

The variables we've been talking about so far have been what are called random variables, in that they all have a theoretical probability distribution. In a sense, a random variable is a set.
There are two kinds of random variables: *discrete* and *continuous*.

- Discrete variables can take on any one of several separate, mutually-exclusive values.
  - Either *finite* or *countably infinite*.
  - e.g. Congressperson's ideology score $\{0, 1, 2, 3..., 100\}$.
- Continuous can take on ANY value in its range.
  - Uncountably infinite number of elements
  - E.g. time, or temperature, or $Y \in [0, 1]$.
  - Most continuous variables are measured discretely.

2.1.1 Probability

The basic probability of an event is equal to:

$$Pr(\text{Event}) = \frac{\text{The number of times the } \textit{event} \text{ can or could occur}}{\text{The number of times } \textit{any event} \text{ can or could occur}}$$

From this definition, there are several important *characteristics* of probabilities that you'll need to remember.

1. Probabilities range between zero and one.

2. The sum of probabilities for all events always equals one.

3. The **Addition Rule**:

   The probability of obtaining *any one* of several independent, mutually exclusive events is equal to the *sum* of the probabilities for those events.

   - That is, $Pr(A \text{ or } B) = Pr(A) + Pr(B)$.
   - Think about this rule in terms of the word "or"...

4. The **Multiplication Rule**:

   The probability of obtaining a *combination* of independent, mutually exclusive events is equal to the *product* of their separate probabilities.

   - That is, $Pr(A \text{ and } B) = Pr(A) \times Pr(B)$.
   - Think about this rule in terms of the word "and"...

2.2 Probability Distributions

Now that we know these basic rules, we can start talking about probability distributions in a bit more detail.

2.2.1 The Probability Density Function (PDF)

If X is some random variable, then the probability that $X =$ some value x in the range of X defines the *probability density function* (PDF). That is, the PDF is the function that maps the values of X to some associated probability of their occurrence.

For *discrete* random variables, we write

$$f(X) = Pr(X = x) \, \forall \, x$$

For a *continuous* probability density function $f(X)$, the probability associated with any given point is zero. Instead, we consider the probability that X takes on a value in some range from (say) A to B:

$$Pr(A \leq X \leq B) = \int_A^B f(X) \, dX$$

where $f(X)$ is the probability density function for X. Note that the rules of probability above mean that, for a continuous PDF $f(X)$, $\int_{-\infty}^{\infty} f(X) \, dX = 1.0$.

2.2.2 The Cumulative Distribution Function (CDF)

If X is at least ordinal, we may be interested in the probability that X will take on a value (say) less than or equal to than some value (say, k) in its range.

In the discrete case, this is equal to:

$$\begin{aligned} Pr(X \leq k) &= \sum_{x \leq k} Pr(X = x) \\ &= 1 - \sum_{x > k} Pr(X = x) \end{aligned}$$

This is the *cumulative distribution function* (CDF), which is sometimes written using capital letters: $f(X) \rightarrow F(X)$.

For continuous variables, we want to know the probability that $X \leq k$, i.e.

$$Pr(X \leq k) = \int_{-\infty}^{k} f(X)\, dX$$

These expositions tell us something:

- *The first derivative of the CDF is the PDF*, and
- *the integral of the PDF is the CDF...*

Moreover, if we write the CDF as $F(X)$, note that:

- $0 \leq F(X) \leq 1$.
- $F(-\infty) = 0$.
- $F(\infty) = 1$.