

## Appendix to chapter 1 outline, Econometrics

### 1 The use of summation operators

Suppose we want to add some numbers. Suppose that there are  $N$  of those numbers. We could write that our sum =  $X_1 + X_2 + X_3 + \dots + X_N$ , where the  $\dots$  means to continue the sequence that we have started until we reach  $X_N$ . However, if  $N = 1000$ , then we would have to write out 1000 numbers. That is very time consuming and unimportant. So we use  $\sum$  as a symbol for the “summation of” a group of numbers. To rewrite  $X_1 + X_2 + X_3 + \dots + X_N$  using the  $\sum$  symbol, we write:

$\sum_{i=1}^N X_i$  – this means we add up all the  $X$ s, starting from  $X_1$  and ending at  $X_N$ .

Below are some rules for using summation operators. Note that this is basically the same stuff that is in the book, but I have added some comments and a little more detail in some places.

#### 1.1 Rule 1

The summation of a constant  $k$  times a variable is equal to the constant times the summation of that variable.

$$\sum_{i=1}^N kX_i = k \sum_{i=1}^N X_i$$

If we were to expand the left-hand side (I will use LHS to abbreviate left-hand side throughout the course, and RHS to abbreviate right-hand side) of the equation, we would get  $kX_1 + kX_2 + kX_3 + \dots + kX_N$ . Note that we can factor out a  $k$  since it appears in each term being summed. This gives us  $k(X_1 + X_2 + X_3 + \dots + X_N)$ . Since  $X_1 + X_2 + X_3 + \dots + X_N = \sum_{i=1}^N X_i$ , we just substitute in and get  $k \sum_{i=1}^N X_i$ .

#### 1.2 Rule 2

The summation of the sum of observations of two variables is equal to the sum of their summations.

$$\sum_{i=1}^N (X_i + Y_i) = \sum_{i=1}^N X_i + \sum_{i=1}^N Y_i$$

If we expand the LHS, we get  $X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3 + \dots + X_N + Y_N$ . We can just use the commutative property of addition and regroup terms as

$(X_1+X_2+X_3+\dots+X_N)+(Y_1+Y_2+Y_3+\dots+Y_N)$ . Since  $X_1+X_2+X_3+\dots+X_N = \sum_{i=1}^N X_i$  and  $Y_1+Y_2+Y_3+\dots+Y_N = \sum_{i=1}^N Y_i$ , we just substitute in to get the RHS.

### 1.3 Rule 3

The summation of a constant over N observations equals the product of the constant and N.

$$\sum_{i=1}^N k = kN$$

If we expand the LHS, all we do is add k N times. It's easy to see if we use  $N = 3$ . Then the LHS is  $k + k + k$ . This equals  $3k$ . This will work for any combination of k and N, and we get the general form on the RHS.

### 1.4 Definition 1

Now that we have 3 rules, we can define the arithmetic mean or average of a variable. The mean (I will use mean to mean arithmetic mean – if we ever need anything like the geometric mean or harmonic mean then I will explicitly state the “geometric mean” or “harmonic mean”) is defined as:

$$\bar{X} = \frac{\sum_{i=1}^N X_i}{N} = \frac{1}{N} \sum_{i=1}^N X_i$$

The concept should not be new to you, although the notation might. It is just the simple average of the numbers – add the observations ( $\sum_{i=1}^N X_i$ ) and divide by the number of observations (N). When you see the symbol  $\bar{X}$ , this represents the mean. You can call this “the mean of X” (which it actually is) or you can call it “X bar”. Most people know what you mean if you say “X bar”.

### 1.5 Rule 4

The summation of the deviations of observations on X about its mean is zero.

$$\sum_{i=1}^N (X_i - \bar{X}) = 0$$

This is a fairly easy proof. First, use rule 2 to get:

$$\sum_{i=1}^N (X_i - \bar{X}) = \sum_{i=1}^N X_i - \sum_{i=1}^N \bar{X}$$

Then, use rule 3 to get:

$$\sum_{i=1}^N X_i - \sum_{i=1}^N \bar{X} = \sum_{i=1}^N X_i - N\bar{X}$$

Note that we can use rule 3 because  $\bar{X}$  is a constant for the sample we have drawn. If we draw another sample of the same variable we may get a different  $\bar{X}$ , but for the one we have drawn it is a constant (no matter how many times we add up those  $N$  observations and divide by  $N$  we will get the same number). Now, use definition 1 to get:

$$\sum_{i=1}^N X_i - N\bar{X} = \sum_{i=1}^N X_i - N \frac{1}{N} \sum_{i=1}^N X_i$$

We know that  $N \frac{1}{N} = 1$ , so  $N \frac{1}{N} \sum_{i=1}^N X_i = \sum_{i=1}^N X_i$ . We now have:

$$\sum_{i=1}^N X_i - \sum_{i=1}^N X_i, \text{ which is equal to zero.}$$

## 1.6 Definition 2

We define the variance of a variable  $X$  to be:

$$\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$$

As for what a variance is in words, think about it this way. We measure deviations from the mean as  $X_i - \bar{X}$ . This tells us how far an observation is from the mean. What we would like to know is “on average how far an observation from a sample is from the mean”. If you think about this logically, it would suggest summing up all the deviations and then dividing by the number of observations. Mathematically, it would look like:

$$\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})$$

But there is a problem with using the above formula for measuring “how far an observation is from the sample mean on average”. The problem is Rule 4,  $\sum_{i=1}^N (X_i - \bar{X}) = 0$ . For any sample we would have, if we used  $\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})$  to measure “how far an observation is from the sample mean on average” we would always get back zero!!! That means that  $\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})$  is a worthless measure.

When we use  $\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2$  (the “real” formula for variance), we don’t have the problem of the squared deviations summing to zero, unless every single point is equal to the mean, in which case we would have “zero variance”, because all the points are the same. What squaring the deviations does (besides making all the terms being summed together positive) is penalize large deviations from the mean (this is discussed a little in the chapter 1 outline). Hopefully you have a better understanding of what variance measures and why we use the formula we use now.

### 1.7 Definition 3

We define the covariance of two variables,  $X$  and  $Y$  to be:

$$\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$$

With covariance, we are trying to measure “on average how closely (in a linear fashion)  $X$  and  $Y$  are related”. If we have a positive covariance, it means that, on average,  $X$  and  $Y$  both tend to be above their means at the same time and below their means at the same time. If we have a negative covariance, this means that, on average,  $X$  tends to be above its mean when  $Y$  is below its mean, and vice versa.

\*\*\*Note: Two variables may have a perfect NONLINEAR relationship, but a zero covariance. This does not mean the variables are unrelated, but it does suggest that they are not related in any meaningful linear fashion.\*\*\*

### 1.8 Rule 5

The covariance between two variables  $X$  and  $Y$  is equal to the mean of the products of observations on  $X$  and  $Y$  minus the product of their means.

$$\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X} \bar{Y}$$

Once again, the proof is fairly simple. First, expand the product of the terms on the LHS. We get:

$$\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{N} \sum_{i=1}^N (X_i Y_i - \bar{X} Y_i + X_i \bar{Y} + \bar{X} \bar{Y})$$

Now, use rule 2 to get:

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N (X_i Y_i - \bar{X} Y_i + X_i \bar{Y} + \bar{X} \bar{Y}) &= \frac{1}{N} \sum_{i=1}^N X_i Y_i - \frac{1}{N} \sum_{i=1}^N \bar{X} Y_i + \frac{1}{N} \sum_{i=1}^N X_i \bar{Y} + \\ &\frac{1}{N} \sum_{i=1}^N \bar{X} \bar{Y} \end{aligned}$$

Notice that the first term on the RHS is the first term on the RHS of what we are trying to prove, so we want to leave that alone. Let's focus on the middle

two terms,  $-\frac{1}{N} \sum_{i=1}^N \bar{X} Y_i + \frac{1}{N} \sum_{i=1}^N X_i \bar{Y}$ . Using rule 1 and definition 1, we get

$$-\frac{1}{N} \sum_{i=1}^N \bar{X} Y_i = -\bar{X} \frac{1}{N} \sum_{i=1}^N Y_i = -\bar{X} \bar{Y} \text{ and}$$

$$-\frac{1}{N} \sum_{i=1}^N X_i \bar{Y} = -\bar{Y} \frac{1}{N} \sum_{i=1}^N X_i = -\bar{Y} \bar{X}$$

So now, our full equation looks like:

$$\frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X} \bar{Y} + \bar{Y} \bar{X} + \frac{1}{N} \sum_{i=1}^N \bar{X} \bar{Y} = \frac{1}{N} \sum_{i=1}^N X_i Y_i - 2\bar{X} \bar{Y} + \frac{1}{N} \sum_{i=1}^N \bar{X} \bar{Y}$$

Now focus on the last term,  $\frac{1}{N} \sum_{i=1}^N \bar{X}\bar{Y}$ . Since  $\bar{X}$  and  $\bar{Y}$  are both constants,  $\bar{X}\bar{Y}$  is a constant. Using rule 3, we know  $\sum_{i=1}^N \bar{X}\bar{Y} = N\bar{X}\bar{Y}$ . Substituting back in, we get that  $\frac{1}{N} \sum_{i=1}^N \bar{X}\bar{Y} = \frac{1}{N} N\bar{X}\bar{Y} = \bar{X}\bar{Y}$ . Now, substituting back in to our main equation we get:

$$\frac{1}{N} \sum_{i=1}^N X_i Y_i - 2\bar{X}\bar{Y} + \frac{1}{N} \sum_{i=1}^N \bar{X}\bar{Y} = \frac{1}{N} \sum_{i=1}^N X_i Y_i - 2\bar{X}\bar{Y} + \bar{X}\bar{Y} = \frac{1}{N} \sum_{i=1}^N X_i Y_i - \bar{X}\bar{Y},$$

which completes our proof.

## 1.9 Rule 6

The variance of X is equal to the mean of the squares of observations of X minus its mean squared.

$$\frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})^2 = \frac{1}{N} \sum_{i=1}^N X_i^2 - (\bar{X})^2$$

The proof of this rule is very similar to the one for rule 5. We just replace  $(Y_i - \bar{Y})$  with  $(X_i - \bar{X})$  and follow the steps.

## 1.10 Rules 7 and 8

Rules 7 and 8 are double summation rules. We will use them on occasion, and I will list them here. I will explain them when necessary. Look at the book for more information, pg. 16.

### 1.10.1 Rule 7

$$\sum_{i=1}^N \sum_{j=1}^N (X_i Y_j) = \sum_{i=1}^N (X_i) \sum_{j=1}^N (Y_j)$$

Notice that since  $X$  is only indexed by  $i$  we can pull it out of the summation over the index  $j$ . We can pull  $Y$  out of the summation over the index  $i$  since it is only indexed by  $j$ .

### 1.10.2 Rule 8

$$\sum_{i=1}^N \sum_{j=1}^N (X_{ij} + Y_{ij}) = \sum_{i=1}^N \sum_{j=1}^N (X_{ij}) + \sum_{i=1}^N \sum_{j=1}^N (Y_{ij})$$

This is just the commutative rule for addition in use again.