

1. To show $Cov(XY) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) = \left[\frac{1}{N} \sum_{i=1}^N X_i Y_i \right] - \bar{X} \bar{Y}$

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

$$Cov(XY) = \frac{1}{N} \left[\sum X_i Y_i - \bar{X} \sum Y_i - \bar{Y} \sum X_i + N \bar{X} \bar{Y} \right]$$

$$Cov(XY) = \frac{1}{N} \left[\sum X_i Y_i - \bar{N} \bar{X} \bar{Y} - N \bar{X} \bar{Y} + N \bar{X} \bar{Y} \right]$$

$$Cov(XY) = \left[\frac{1}{N} \sum_{i=1}^N X_i Y_i \right] - \bar{X} \bar{Y}$$

First find \bar{X} and \bar{Y}

$$\bar{X} = \frac{1}{5}(11+14+12+16+12) = \frac{65}{5} = 13$$

$$\bar{Y} = \frac{1}{5}(1+6+8+10+5) = \frac{30}{5} = 6$$

I	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i Y_i)$
1	-2	-5	11
2	1	0	84
3	-1	2	96
4	3	4	160
5	-1	-1	60

So using either formula

$$Cov(XY) = \frac{1}{5}(10+0-2+12+1) = \frac{21}{5} = 4.2$$

or

$$Cov(XY) = \left[\frac{1}{N} \sum_{i=1}^N X_i Y_i \right] - \bar{X} \bar{Y} = \frac{1}{5}(11+84+96+160+60) - (13*6)$$

$$= 82.2 - 78$$

$$= 4.2$$

Stata commands to obtain sample variance and covariance

```
. list
```

	age	yearsed	tenure
1.	18	11	1
2.	29	14	6
3.	33	12	8
4.	35	16	10
5.	45	12	5

```
. su yearsed, detail
```

yearsed			
Percentiles		Smallest	
1%	11	11	
5%	11	12	
10%	11	12	Obs 5
25%	12	14	Sum of Wgt. 5
50%	12		Mean 13
		Largest	Std. Dev. 2
75%	14	12	
90%	16	12	Variance 4
95%	16	14	Skewness .6288941
99%	16	16	Kurtosis 1.953125

```
. di (4*4)/5
```

```
3.2
```

```
. corr yearsed tenure, cov  
(obs=5)
```

	yearsed	tenure
yearsed	4	
tenure	5.25	11.5

```
. di (5.25*4)/5
```

```
4.2
```

2. To show

$$\text{Var}(X) = \frac{1}{N} \sum_i (X_i - \bar{X})^2 = \frac{1}{N} \sum_i (X_i)^2 - \bar{X}^2$$

$$\begin{aligned} \frac{1}{N} \sum_i (X_i - \bar{X})^2 &= \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(X_i - \bar{X}) = \frac{1}{N} \sum_{i=1}^N (X_i^2 - \bar{X} X_i - \bar{X} X_i + \bar{X}^2) \\ &= \frac{1}{N} \sum_{i=1}^N (X_i^2 - 2\bar{X} X_i + \bar{X}^2) \end{aligned}$$

Using $\sum_{i=1}^N X_i = N \bar{X}$ and separating terms in brackets

$$\begin{aligned} &= \frac{1}{N} \sum_{i=1}^N (X_i^2) - \frac{2N\bar{X}}{N} + \frac{N\bar{X}^2}{N} \\ &= \frac{1}{N} \sum_i (X_i)^2 - \bar{X}^2 \end{aligned}$$

So to find Var(X)

i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	(X_i^2)
1	-2	4	121
2	1	1	196
3	-1	1	144
4	3	9	256
5	-1	1	144

Either

$$\text{Var}(X) = \frac{1}{N} \sum (X_i - \bar{X})^2 = \frac{1}{5}(4+1+1+9+1) = \frac{16}{5} = 3.2$$

or

$$\text{Var}(X) = \frac{1}{N} \sum_i (X_i)^2 - \bar{X}^2 = \frac{1}{5}(121+196+144+256+144) - 169 = \frac{861}{5} - 169$$

$$= 172.2 - 169 = 3.2$$

Note that if the X data are multiplied by 10

$$\bar{X} = \frac{1}{5}(110 + 140 + 120 + 160 + 120) = \frac{650}{5} = 130$$

then the mean is also multiplied by 10

and the variance

$$\begin{aligned} \text{Var}(X) &= \frac{1}{N} \sum_i (X_i)^2 - \bar{X}^2 = \frac{1}{5}(12100 + 19600 + 14400 + 25600 + 14400) - 16900 = \frac{86100}{5} - 16900 \\ &= 320 \end{aligned}$$

the variance is therefore multiplied by 100 if the data are multiplied by 10

[and in general $\text{Var}(aX) = a^2\text{Var}(X)$ if a is a constant]

Similarly the rules on covariances imply that

$$\text{Cov}(aX, Y) = a\text{Cov}(XY) \quad (\text{see question 3})$$

So

$$\begin{aligned} \text{Cov}(XY) &= \left[\frac{1}{N} \sum_{i=1}^N X_i Y_i \right] - \bar{X} \bar{Y} = \frac{1}{5}(110 + 840 + 960 + 1600 + 600) - (130 * 6) \\ &= 822 - 780 \\ &= 42 \end{aligned}$$

so the covariance is multiplied by 10 when the X data are multiplied by 10

These results help illustrate that neither the variance nor the covariance are scale invariant – their values will depend on the units of measurement of the variables

3. If $Y = A + B$, show that

$$\text{Cov}(X, Y) = \text{Cov}(X, A) + \text{Cov}(X, B)$$

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

and since $Y = A + B$ then $\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{\sum_{i=1}^N A_i + B_i}{N} = \frac{\sum_{i=1}^N A_i}{N} + \frac{\sum_{i=1}^N B_i}{N} = \bar{A} + \bar{B}$

$$\text{So } \text{Cov}(XY) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(A_i + B_i - \bar{A} - \bar{B})$$

$$= \frac{1}{N} \left(\sum_{i=1}^N (X_i - \bar{X})(A_i - \bar{A}) + \sum_{i=1}^N (X_i - \bar{X})(B_i - \bar{B}) \right)$$

$$= \text{Cov}(XA) + \text{Cov}(XB)$$

It follows that

$$\text{Var}(Y) = \text{Var}(A) + \text{Var}(B) + 2\text{Cov}(A, B)$$

Since

$$\text{Var}(Y) = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})(Y_i - \bar{Y}) = \text{Cov}(YY)$$

$$= \text{Cov}(Y, (A+B)) = \text{Cov}[(A+B), (A+B)] = \text{Cov}[(A+B), A] + \text{Cov}[(A+B), B]$$

(since $\text{Cov}(XY) = \text{Cov}(XA) + \text{Cov}(XB)$ if $Y = A + B$)

$$= \text{Cov}(A, A) + \text{Cov}(B, A) + \text{Cov}(A, B) + \text{Cov}(B, B)$$

$$= \text{Var}(A) + \text{Var}(B) + 2\text{Cov}(A, B)$$

ii) To show $\text{Cov}(XY) = a\text{Cov}(XZ)$ if $Y = aZ$

$$\text{Cov}(XY) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(aZ_i - a\bar{Z})$$

(since $\bar{Y} = \frac{\sum_{i=1}^N y_i}{N} = \frac{\sum_{i=1}^N aZ_i}{N} = a \frac{\sum_{i=1}^N Z_i}{N} = a\bar{Z}$)

$$\text{Cov}(XY) = \frac{a}{N} \sum_{i=1}^N (X_i - \bar{X})(Z_i - \bar{Z})$$

So $\text{Cov}(XY) = a\text{Cov}(XZ)$

iii) To show $\text{Cov}(X, Y) = 0$ if $y = a$ (constant)

$$\bar{a} = \frac{\sum_{i=1}^N a_i}{N} = \frac{Na}{N} = a$$

So $(a_i - \bar{a}) = 0$ for all i

$$\text{So } \text{Cov}(XY) = \frac{1}{N} \sum_{i=1}^N (X_i - \bar{X})(a_i - \bar{a}) = 0$$

To show

$$\text{Cov}(\hat{Y}, \hat{u}) = 0$$

$$\text{Let } \hat{y} = \hat{b}_0 + \hat{b}_1 X$$

$$\text{So } \text{Cov}(\hat{Y}, \hat{u}) = \text{Cov}(\hat{y} = \hat{b}_0 + \hat{b}_1 X, \hat{u}) = \text{Cov}(\hat{b}_0 \hat{u}) + \text{Cov}(\hat{b}_1 X \hat{u})$$

Since \hat{b}_0 is a constant then $\text{Cov}(\hat{b}_0 \hat{u}) = 0$

And since \hat{b}_1 is a constant then $\text{Cov}(\hat{b}_1 X, \hat{u}) = \hat{b}_1 \text{Cov}(X, \hat{u})$

$$\text{Let } \hat{u} = y - \hat{y} = y - \hat{b}_0 - \hat{b}_1 X$$

$$\text{So } \hat{b}_1 \text{Cov}(X, \hat{u}) = \hat{b}_1 \text{Cov}[X, (y - \hat{b}_0 - \hat{b}_1 X)]$$

$$= \hat{b}_1 [\text{Cov}(X, y) - \text{Cov}(X, \hat{b}_0) - \text{Cov}(X, \hat{b}_1 X)]$$

$$= \hat{b}_1 [\text{Cov}(X, y) - \text{Cov}(X, \hat{b}_0) - \hat{b}_1 \text{Var}(X)]$$

from above $\text{Cov}(\hat{b}_0 \hat{u}) = 0$

and we know the OLS formula $\hat{b}_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}$

So

$$\text{Cov}(\hat{y}, \hat{u}) = \hat{b}_1 [\text{Cov}(X, Y) - \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \text{Var}(X)] = 0$$

4. Correlation coefficient given by

$$r_{xy} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}}$$

from answers to earlier questions it follows that

$$r_{xy} = \frac{4.2}{\sqrt{3.2 * 9.2}} = 0.77$$

so X (years of education) and Y (job tenure) are positively related

Note that if the X data are multiplied by 10 then

$$r_{xy} = \frac{42}{\sqrt{320 * 9.2}} = 0.77$$

so correlation coefficient (unlike the variance and covariance) is unchanged when the data are re-scaled

- said to be scale invariant

5. Given $Y = 4000 + 0.7X$

this is a simple linear equation which traces out a straight line with an intercept (= 4000) and a slope (=0.7)

So for every £1 of before tax income after tax income rises by 70 pence (slope = dY/dX so $dY = \text{slope} * dX$)

Follows that the mean $E(Y) = E[4000 + 0.7X]$

Since 4000 is a constant its expected value is always the same, $E(4000) = 4000$

Since X is a random variable it fluctuates around an average (mean) value

$$E(0.7X) = 0.7E(X) = 0.7\mu_x$$

$$\text{So } E(Y) = 4000 + 0.7 \mu_x$$

Follows that the variance of Y given by

$$E[(Y - \mu_y)^2] = E[(4000 + 0.7X - 4000 - 0.7\mu_x)^2]$$

$$E[(Y - \mu_y)^2] = E[(0.7(X - \mu_x))^2] = 0.49E[(X - \mu_x)^2]$$

so

$$\text{Var}(y) = 0.49\text{Var}(X)$$

Since standard deviation is square root of variance

$$\text{s.d.}(y) = 0.7\text{s.d.}(X)$$

(standard deviation of after tax income is 70% of standard deviation in before-tax income)