Dynamics II: Statistics, 4. June 2007

1) Mean, Standart deviation etc.:

a) Calculate for the three data sets (0, 0, 14, 14), (0, 6, 8, 14) and (6, 6, 8, 8), means, standard deviations, sample standard deviations!

b) The Pearson correlation coefficient is then the best estimate of the correlation of X and Y. Calculate this coefficient for the data in a)!

2) Sample correlation coefficient and explained variance:

a) Show: The square of the sample correlation coefficient, also known as the coefficient of determination, is the fraction of the variance in y_i that is accounted for by a linear fit of x_i to y_i :

$$r_{xy}^2 = 1 - \frac{S_{y|x}^2}{S_y^2} \tag{1}$$

Use:

$$S_{y|x}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \tag{2}$$

which is the square of the error of a linear regression of x_i on y_i by the equation $y = a + bx$.

b) The linear equation that best describes the relationship between x and y can be found by linear regression. This equation can be used to "predict" the value of one measurement from knowledge of the other. That is, for each value of x the equation calculates a value which is the best estimate of the values of y corresponding the specific value. We denote this predicted variable by y. Any value of Y can therefore be defined as the sum of y' and the difference between y and y':

$$y = y' + \ (y - y') \tag{3}$$

Use

$$S_y^2 = S_{y'}^2 + S_{y|x}^2 \tag{4}$$

to derive an expression for r_{xy}^2 !

c) with b). The square of r is conventionally used as a measure of the association between x and y. For example, if the coefficient is 0.90, then ??? % of the variance of y can be "accounted for" by changes in x and the linear relationship between x and y.

3) Correlation and significance :

Using the following piece of code in R, evaluate corr, t-value, degrees of freedom, significance of the correlation for the time series a and b.

```
a<- 1:1000
noise<- rnorm(1000)*500
b<- a + noise
plot(a,b)
cor(a,b)
cor.test(a,b)
```