

APPLIED QUANTITATIVE REASONING 601

COMPUTER LAB#4: Linear Regression

We perform a Monte Carlo simulation of the linear regression model.

- N is number of data points
- X is the independent (explanatory) variable
- Y is the dependent variable
- u is the stochastic disturbance,
- a is the Y intercept and b is the slope (population values).
- a and b are sample estimates of the parameters a and b respectively

We create a vector of random numbers from a normal distribution with mean $\mu=0$ and standard deviation $\sigma=6$.

Assign the mean μ value: $\mu := 0$

Assign the standard deviation σ value: $\sigma := 6$

Normally distributed random deviates (errors) are generated by using the Mathcad function $\text{rnorm}(N, \mu, \sigma)$, where N is the number of data points, μ is the population mean and σ is the population standard deviation:

$$u := \text{rnorm}(N, \mu, \sigma)$$

We assign values to the population coefficients α and β are:

$$\alpha := 5 \quad \beta := 2$$

The sample size N is defined by using the global assignment operator.

$$i := 0..N-1$$

The explanatory variable X and the dependent quantity Y are defined:

$$X_i := i \quad Y := \beta \cdot X + \alpha + u$$

a and b, the sample estimates for α and β , are obtained by using the MathCAD functions $\text{intercept}(X, Y)$ and $\text{slope}(X, Y)$.

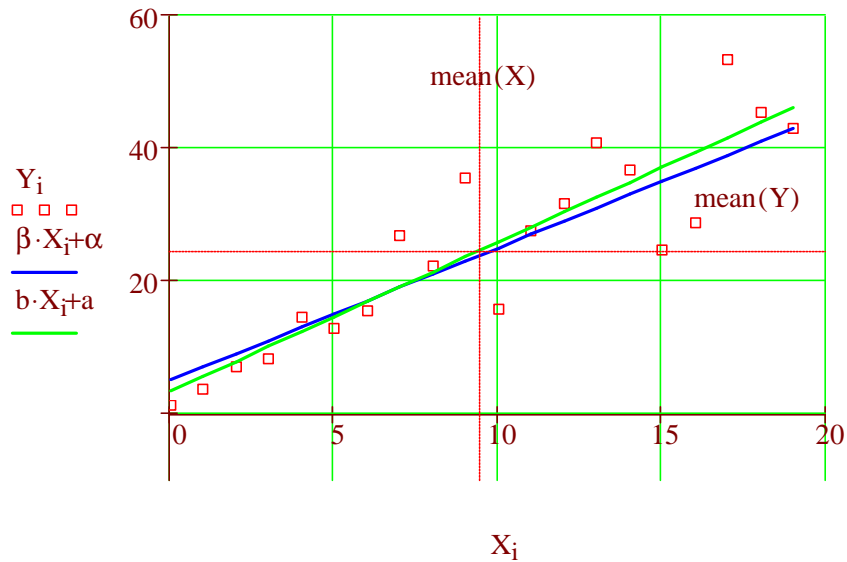
$$a := \text{intercept}(X, Y)$$

$$b := \text{slope}(X, Y)$$

Now we are ready for the simulation of a sample regression line from a population whose coefficients are known ($a=5$ and $b=2$).

Enter the number of random deviates N (here, 20, corresponding to the sample size). Click on N=20 below, then click F9 (compute). Type the estimates a and b in two arrays (vectors, prepared below). Repeat the experiment 15 times. Each time record the estimates.

$N \equiv 20$



$$a = 3.351$$

$$b = 2.250$$

The Pearson's correlation coefficient r is a measure of the linear correlation between the X and Y values.

A value of r close to 0 signifies a low level of correlation (in other words the two variables are not linearly dependent).

$$r := \text{corr}(X, Y)$$

A value of r close to +1 or -1 signifies a high degree of linear correlation.

$$r = 0.897$$

The correlation coefficient r has the same as the slope β . In MathCAD, the Pearson coefficient is given by the built-in function: $\text{corr}(X, Y)$.

Watch what happens to the r value as you click on F9 to simulate the resampling.

Here are my estimates for a set of 15 runs (equivalent to sampling 20 pairs of values X, Y 15 times from a population with known parameters values $a=5$ and $b=2$).

	2.058	2.258
	1.570	2.230
	1.959	2.194
	6.652	1.721
	-1.104	2.429
	6.036	1.881
	1.929	2.357
a :=	5.260	b := 1.685
	6.184	1.865
	8.273	1.641
	4.934	2.163
	2.722	2.168
	3.496	2.015
	7.503	1.641
	6.178	1.745

Next we calculate the means of our 15 pairs of coefficient estimates a and b (stored in the two vectors above) and compare them to the (known) population values: $a = 5$ and $b = 2$.

This is an experimental check on the property of OLS estimates to be unbiased:

$$\text{mean}(a) = 4.243$$

$$\text{mean}(b) = 2.000$$

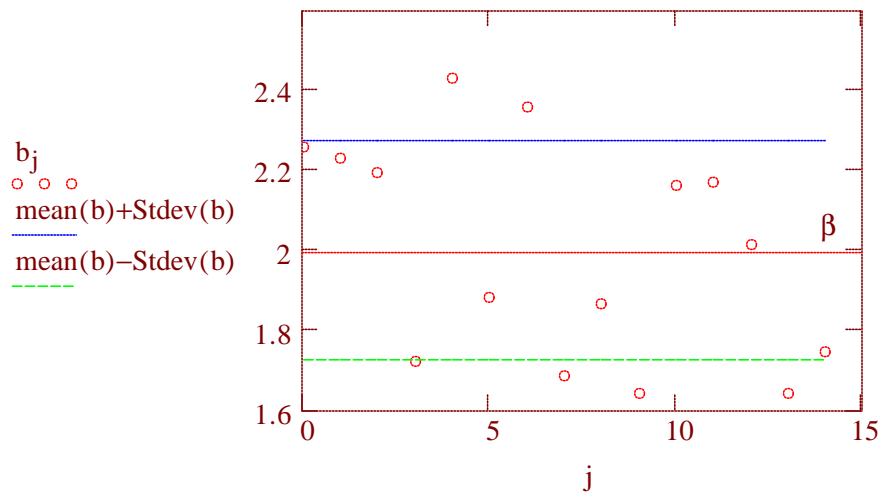
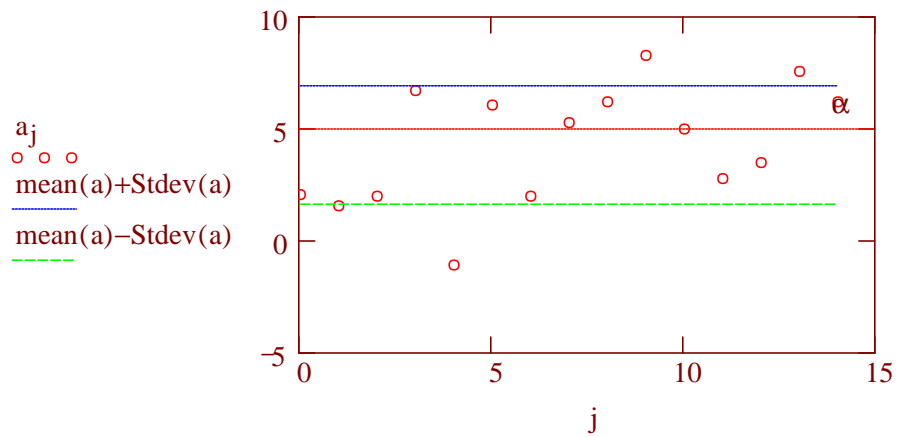
We then compute the standard deviations for those estimates.

$$\text{Stdev}(a) = 2.657$$

$$\text{Stdev}(b) = 0.275$$

$$j := 0..14$$

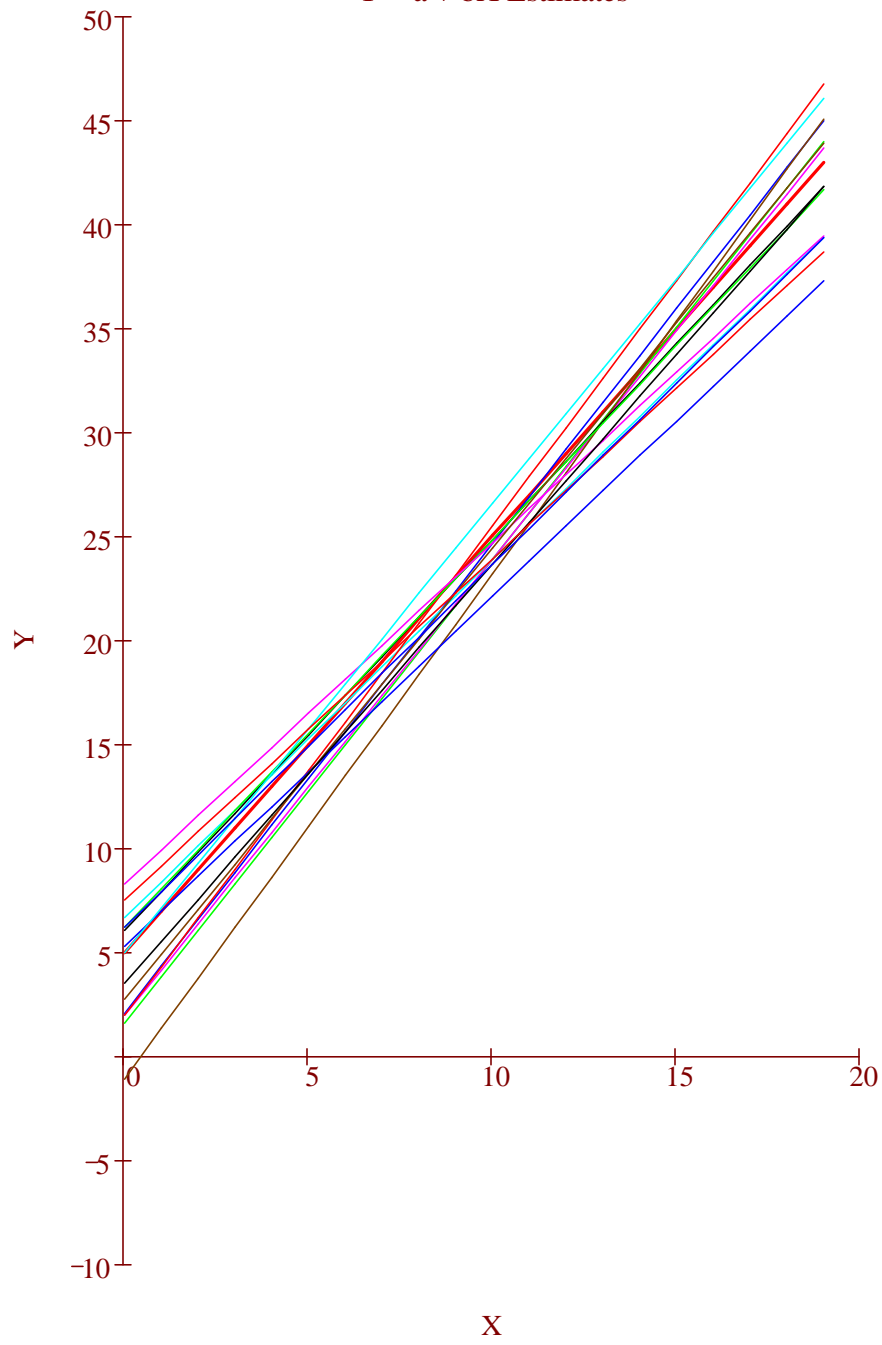
We graph below the intercept estimates a and the slope estimates b from our 15 simulations, and compare them to their respective means and to values within one standard error away from the means:



Now we graph below the regression lines obtained from simulating 15 samples of observation pairs X and Y drawn from a population with known parameters.

Note that each is different and note also the way in which they differ from one another.

$Y = a + bX$ Estimates



FOOD FOR THOUGHT

What have you observed in this simulation?

- As you hit F9 to resample and recompute, watch the sample mean point (meanX, meanY). Is it always on the sample regression line? Why?
- How does an estimated sample regression slope b from a simulated sample compare to the population slope β ?
- What are the means of the estimated intercepts and slopes respectively from 15 simulations? How do they compare to the real population values α and β you set at the beginning of the simulation?

If you feel bold, you can change in this file:

α
 β
N

Try to set $\beta=0$ and watch what happens to the estimated regression line.

- Are the estimated slopes also 0? Why or why not?
- What do your results mean with respect to the t test of slope?

If you feel even bolder, you can (carefully) extend the two arrays to accommodate more than 15 simulations.