VI. ESTIMATION – PART II
LINEAR REGRESSION AND CORRELATION

Linear Regression:

Let *x* and *y* denote two jointly observed numeric variables. This means either that they are both defined for members of the same population, or that they both are parts of the outcome of the same experiment.

Example 1: Let *x* be the height and *y* the weight of individuals from a human population.

Example 2: Let *x* be the amount of fertilizer applied to a plot of cotton seedlings and let *y* be the weight of raw cotton harvested at maturity.

There is a subtle difference between Examples 1 and 2. In Example 1 both variables are observed but neither is controlled by the observer. In Example 2, the experimenter might control the amount of fertilizer as part of the experimental protocol, but otherwise he or she would not have any control over the amount harvested. Example 2 is an example of a *designed experiment* with a random response (the variable *y*). Example 1 is an example of an *observational study*. Despite the difference, the techniques of linear regression, which include least-squares estimation, are used for both designed experiments and observational studies.

Observational studies are likely to occur in education, political science, economics and other social sciences. Designed experiments are found more often in medical and biological science, physical science, and engineering.

Discussion: Think of some exceptions to these generalities. Open some data sets and identify pairs of jointly observed variables *x* and *y*. Would they be better described as results of designed experiments or observational studies?

Linear regression proceeds from certain assumptions:

(*i*) There is a population or distribution of values of *y* for any particular value of *x*.

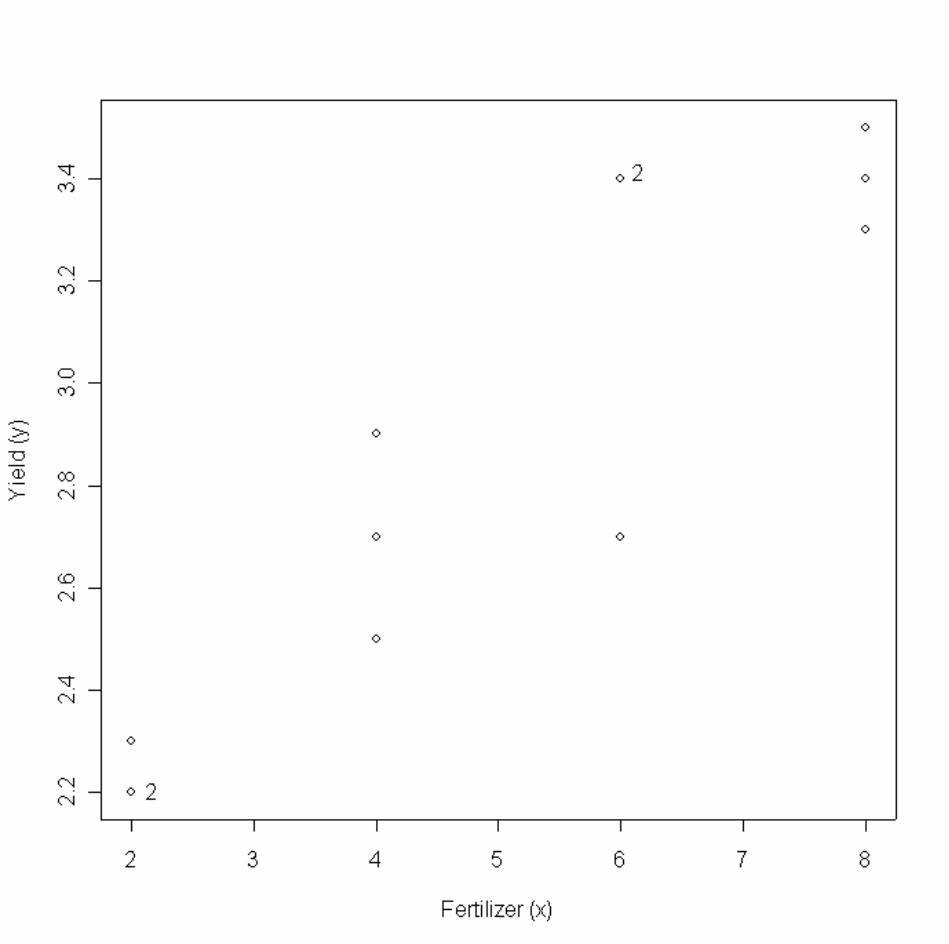
(*ii*) There are constants *a* and *b* so that for any particular value of *x*, the mean of all the corresponding values of *y* is $\mu_y = a + bx$.

(*iii*) The standard deviation σ of the values of *y* corresponding to a value of *x* is the same for all values of *x*.

Assumption (*ii*) is why this technique is called linear regression. It says that the mean value of *y* is a linear function of *x*. If different values of μ_x and corresponding values of μ_y could be plotted on a rectangular coordinate system, the graph would be a straight line with slope *b*. This is called the *theoretical regression line*. The line cannot be plotted because the constants *a* and *b* are unknown. However, they can be estimated by the method of least squares.

To estimate the regression line, we must have data in the form of *n* pairs $(x_1, y_1), \cdots, (x_n, y_{n)}$ of values of *x* and corresponding values of *y*. If these points are plotted on a rectangular coordinate system, they will not lie on a straight line. The reason is that each value of *y* differs from its expected value by some random deviation. The plot of the points (x_i, y_i) is called a scatterplot of the data. Below is a scatterplot of cotton yields (*y*) against fertilizer (*x*).

x	2	2	2	4	4	4	6	6	6	8	8	8
y	2.3	2.2	2.2	2.5	2.9	2.7	3.4	2.7	3.4	3.5	3.4	3.3



The strategy in least-squares estimation is to choose the values of *a* and *b* that minimize the sum of the squared differences between the observed values y_i and the putative expected values $a + bx_i$. In other words, we choose *a* and *b* to make the quantity

$$(y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \cdots + (y_n - a - bx_n)^2$$

as small as possible.

To derive the solution of this problem involves a lengthy argument, which we will skip. The numerical values of *a* and *b* may be obtained with a calculator or with computer software. For small data sets, it is not hard to find the solution by hand, using the following steps.

(*a*) Find the averages $\bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$ and $\bar{y} = \frac{1}{n}(y_1 + \cdots + y_n)$.

(*b*) Find the sample variance $s_x^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]$ of the *x* values.

(*c*) Find the *covariance* $s_{xy} = \frac{1}{n-1}[(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})]$ between the *x* values and the *y* values.

(*d*) The slope *b* is given by $b = \frac{s_{xy}}{s_x^2}$.

(*e*) The intercept *a* is given by $a = \bar{y} - b\bar{x}$.

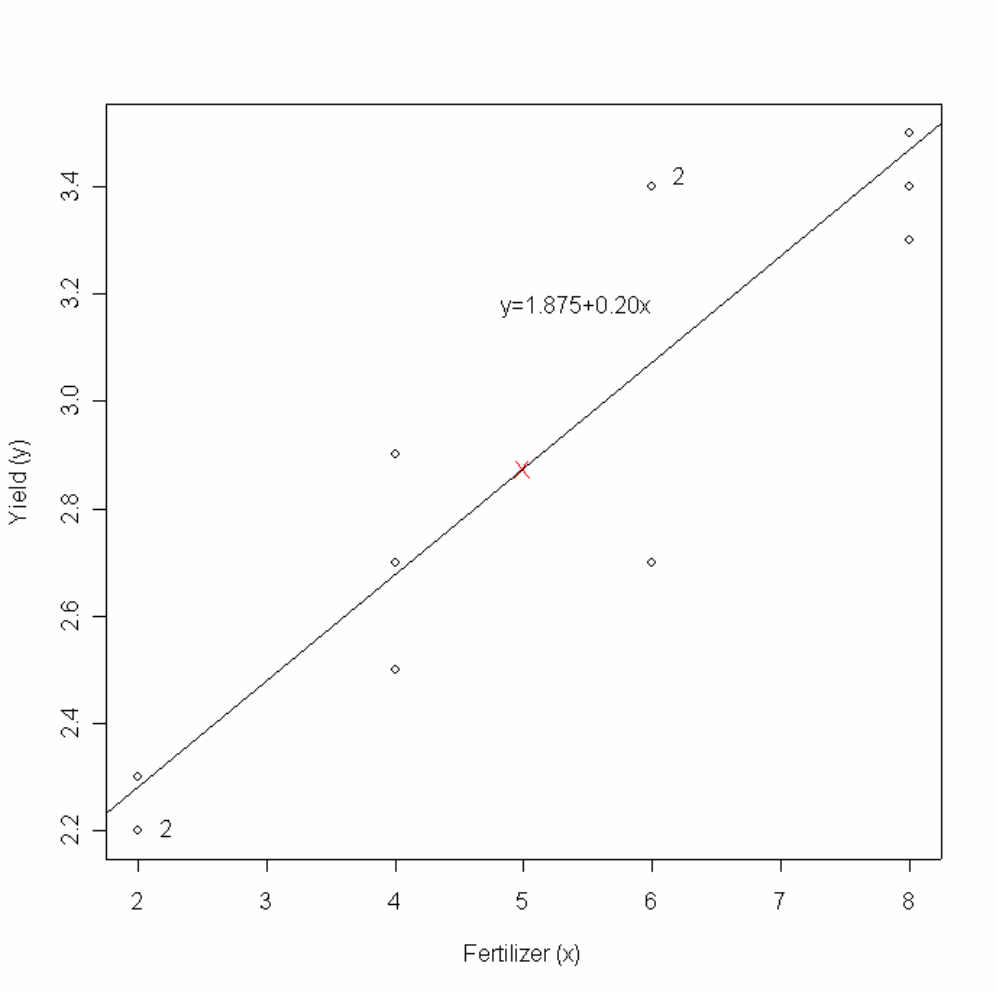
Several remarks are in order. First, the denominators $n - 1$ in (*b*) and (*c*) will cancel in (*d*), so it is not really necessary to include them in the calculations. Second, the slope *b* is usually much more important to the researcher than the intercept *a*. In fact, it is not necessary to calculate *a* at all if it is of no particular interest. This is because the least-squares regression line always passes through the point $(\bar{x}, \bar{y})$. The point-slope form of the equation of a straight line can be used instead of the slope-intercept form. Third, "covariance" is a new term. The equation in (*c*) may be taken as its definition.

Example 6.1:

We will find the equation of the least squares regression line for the cotton yield experiment. The average fertilizer application is $\bar{x} = 5$. The average yield is $\bar{y} = 2.875$. The sample variance of *x* is $s_x^2 = 5.4545$. The covariance between x and *y* is

$$s_{xy} = \frac{1}{11}[(2-5)(2.3-2.875) + (2-5)(2.2-2.875) + \cdots (8-5)(3.3-2.875)] = 1.0818.$$

Thus, the slope of the regression line is $b = \frac{1.0818}{5.4545} = 0.20$, to two-place accuracy. The intercept is $a = 2.875 - 0.20 \times 5 = 1.875$. Using the point-slope form, the equation of the line is $y - \bar{y} = b(x - \bar{x})$, or $y = 2.875 + 0.20(x - 5)$. The plotted line is shown below.



Exercise: Open the data set *nlschools*. One of the variables is *Language*, meaning language test score and another variable is *SocEcoStatus*, for socio-economic status. Select a random sample of 10 students from this data set. Make a scatterplot of *Language* on the vertical scale and *SocEcoStatus* on the horizontal scale. Calculate the least squares regression line for these 10 students. Do the calculations without using any of the higher functions on your calculator. In other words, use it only to do the arithmetic. Next, select a random sample of 50 students and repeat the steps above. This time use the full functionality of your calculator or spreadsheet program.

Correlation:

The *correlation* between two variables *x* and *y* is a measure of how closely associated they are, or how nearly linearly related they are. Correlation is more useful in observational studies than in designed experiments.

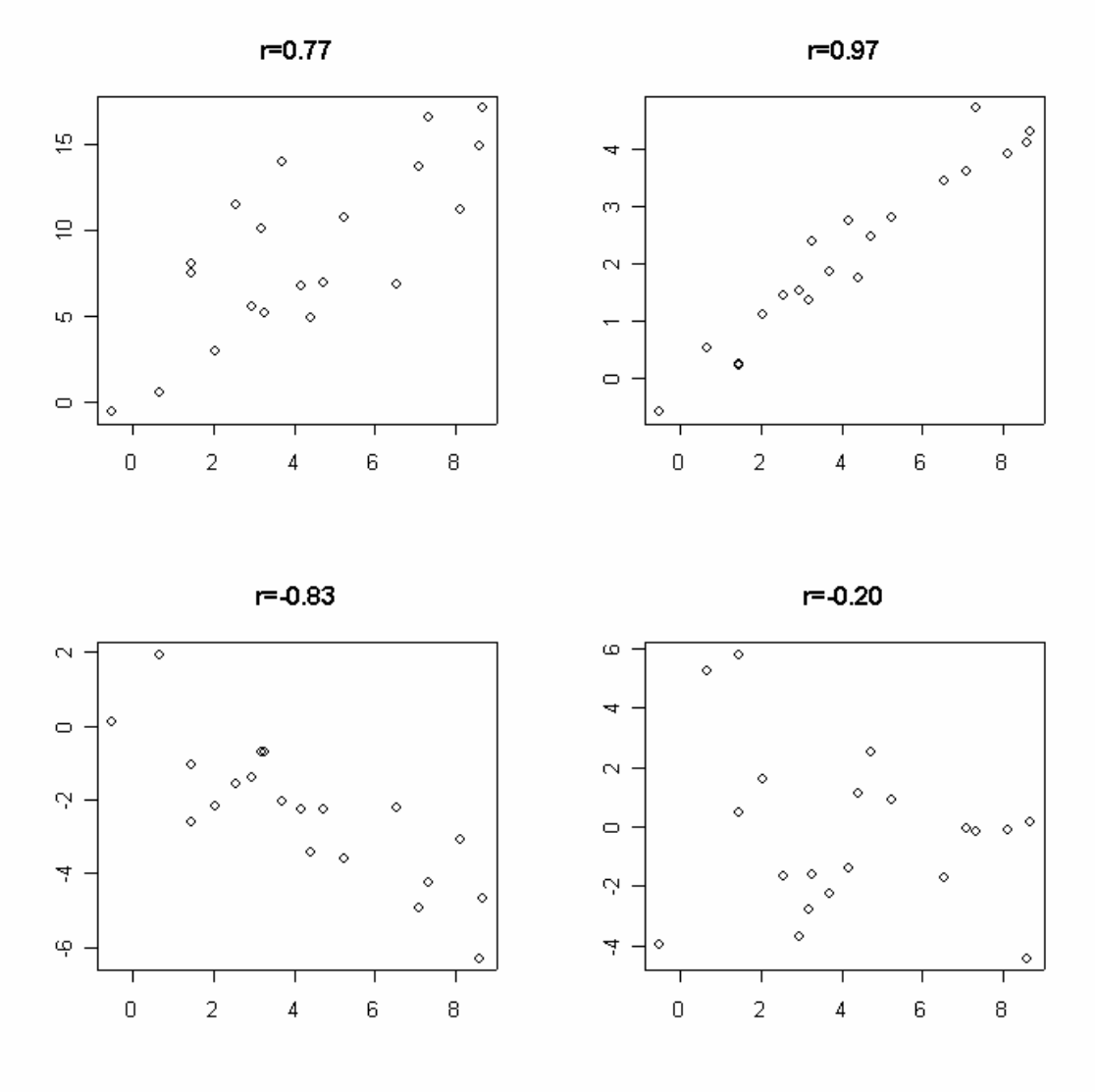
Let $(x_1, y_1), \cdots, (x_n, y_n)$ be joint observations of two variables *x* and *y*. The *correlation* between the values of *x* and the values of *y* is their covariance divided by the product of their sample standard deviations. In symbols,

$$r = \frac{s_{xy}}{s_x s_y}.$$

The correlation is a number between -1 and 1. In fact, it is closely related to the slope of the least-squares regression line.

$$r = b\frac{s_x}{s_y}$$

This shows that *r* and *b* always have the same sign. The regression line slopes upward if $r > 0$ and it slopes downward if $r < 0$. If $r = \pm 1$, the line fits the data (x_i, y_i) perfectly. However, this never happens with real data. If $r = 0$, the variables *x* and *y* are said to be *uncorrelated.* The plots below show several pairs of variables, with different degrees of correlation.



Exercise: Open the sleep data set. Construct scatterplots of the logs of body weight vs. logs of brain weight and body weight vs. total sleep. By comparing them to the examples above, try to guess the correlations between the variables. Then calculate the correlations.