Regression and Correlation

Least squares linear regression Given n pairs of measurements on two variables x and y:

x	$x_1,\ x_2,\ \cdots,\ x_n$
y	$y_1,\ y_2,\ \cdots,\ y_n$

we first plot them to get a rough idea of the relationship (if any) between x and y. Suppose we see a linear relationship and would like to fit a straight line

$$y = a_0 + a_1 x$$

to the data. Our task is to find estimates of a_0 and a_1 such that the line gives a good fit. One way of doing this is by the *method of least squares*. Denote by $\hat{a}_0$ and $\hat{a}_1$ the least squares estimates of a_0 and a_1, respectively. The line

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x$$

is called the *least square linear regression* of y on x. The least squares estimates $\hat{a}_0$ and $\hat{a}_1$ can be computed as follows:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \\
\bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i, \\
S_{xx} &= \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}, \\
\hat{a}_1 &= \frac{S_{xy}}{S_{xx}}, \\
\hat{a}_0 &= \bar{y} - \hat{a}_1\bar{x}.
\end{aligned}$$

Sometimes it is helpful to perform as a table

x	x_1	x_2	$\cdots$	x_n	Σx_i
y	y_1	y_2	$\cdots$	y_n	Σy_i
x^2	x_1^2	x_2^2	$\cdots$	x_n^2	Σx_i^2
xy	x_1y_1	x_2y_2	$\cdots$	x_ny_n	Σx_iy_i
y^2	y_1^2	y_2^2	$\cdots$	y_n^2	Σy_i^2

Correlation coefficient

The most important measure of the degree of correlation between two variables is a quantity called the (product moment) *correlation coefficient*

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}},$$

where

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2.$$

It can be shown that $r \in [-1, +1]$. For $r = +1$, all the observed points lie on a straight line which has a positive slope; for $r = -1$, all the observed points lie on a straight line which has a negative slope; for r near $+1$ (resp. -1), there is a strong positive (resp. negative) linear relationship between two variables. The correlation is significantly different from zero at the α level of significance if

$$|r|\sqrt{\frac{n-2}{1-r^2}} \geq t_{\alpha/2,n-2}.$$