Relationship between two variable quantities - Vary together

Magnitude of dependent variable (y-axis) depends on magnitude of independent variable (x-axis)

Positive
   Both go up or down together

Negative
   One goes up as the other goes down

Correlation: Measure of the degree to which two variables vary together
   Dependent and independent variable are interchangeable, e.g. length and width of leaves, length of forearm and height
   Both measured with error
   Joint distribution
   Bivariate normal distribution (each is normal)
   Not cause and effect

Population correlation coefficient $\rho$
   Measures strength of (linear) relationship
   If two variables are statistically independent, $\rho = 0$
   Estimated by sample correlation coefficient, $r$
   Can be positive or negative, from -1 to +1
   Need probability statement regarding possibility of chance occurrence of $r$

Regression
   Amount of change in dependent variable per unit change in independent variable
   Assumed to be linear
   Independent variable is assumed to be measured without error (fixed)
   Not bivariate normal distribution, $r$ has no meaning
   Can be interpreted as cause and effect

Coefficient of determination, $r^2$
   Square of $r$
   Positive, from 0 to 1
   Proportion of total treatment SS accounted for by regression
   $Y$ vs $X$
   Mathematical model, for given $X_0$, $Y = \alpha + \beta X_0 + \varepsilon$

Coefficient of multiple determination, $R^2$
   $Y$ vs $X_1, X_2, X_3$, etc.

Calculation of coefficient of linear correlation, $r$
or

\[ r' = \frac{\sum (x & y)}{\sqrt{\sum x^2 \sum y^2}} \]

where \( x = X - \bar{X} \) and \( y = Y - \bar{Y} \). Thus \( r = SP/O(SSX)(SSY) \). SP can be either positive or negative.

This is equivalent to:

\[ r' = \frac{\sum X_i Y_i \delta \sum X \sum Y / n}{\sqrt{(\sum X_i^2 \delta \sum X)^2 / n (\sum Y_i^2 \delta \sum Y)^2 / n)} \]

Testing null hypothesis \( H_0: \rho = 0 \) tests whether the variables are independent. The test statistic (with \( n - 2 \) degrees of freedom) is:

\[ t' = \frac{r \sqrt{n \delta^2}}{\sqrt{1 \delta r^2}} \]

Since this involves only \( n \) and \( r \), one can look up \( r \) for the appropriate degrees of freedom in Table A7 (L & H pg 310). This is a two-tailed test. The advantage of calculating the \( t \) statistic is that it can be used to calculate confidence limits.

Calculation of coefficient of determination, \( r^2 \)

\[ r^2 = \frac{(\sum X_i Y_i \delta \sum X \sum Y / n)^2}{(\sum X_i^2 \delta \sum X)^2 / n (\sum Y_i^2 \delta \sum Y)^2 / n)} \]

Degrees of Freedom for \( r \) and \( r^2 \)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>Total</th>
<th>n-1</th>
<th>Regression</th>
<th>1</th>
<th>Error</th>
<th>n-2</th>
</tr>
</thead>
</table>

6.16
**Cause and effect**

Correlation is not sufficient to show cause and effect. eg age and number of grandchildren

Regression is not sufficient to show cause and effect. eg nitrogen and crop response, until research on nitrogen uptake is done it is just as possible that improved growth is due to nitrogen killing nematodes in the soil and not a direct effect

For cause and effect need:
- Variables are related
- Relationship is dose-dependent
- Response is absent in absence of cause
- There is a direct physical method of response, eg uptake, receptor

**Regression Line**

\[ y = a + bx \]

Predicted value of \( Y \) is \( \hat{Y} \). Best estimate of \( \hat{Y} \) from least squares gives

\[ \hat{Y} = \frac{\Sigma xy}{\Sigma x^2} \]

where \( x = X_i - \bar{X} \)

and \( y = Y_i - \bar{Y} \) are deviations from the mean

\( b_{y,x} \) is the regression coefficient of \( Y \) on \( X \)

\[ b = \frac{SP}{SSX} = \frac{\Sigma xy}{\Sigma x^2} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2} \]

For any point on the regression line, the slope is

\[ b = \frac{Y_i - \bar{Y}}{X_i - \bar{X}} \]

then

\[ \hat{Y} - \bar{Y} = b(X - \bar{X}) = bX - b\bar{X} \]

\[ \hat{Y} = \bar{Y} - b\bar{X} = \bar{Y} - b\bar{X} - bX \]

Let \( a = \bar{Y} - b\bar{X} \)

then \( \hat{Y} = a + bX \)

where \( a \) is the intercept and \( b \) is the slope.
For the regression of Y on X: \( b_{y,x} = \frac{SP}{SSX} \)

For the regression on X on Y: \( b_{x,y} = \frac{SP}{SSY} \)

Then \( b_{y,x} \cdot b_{x,y} = \frac{(SP)^2}{(SSX)(SSY)} = r^2 \)

\( r^2 \cdot 100\% \) is the percentage of the variation accounted for by the regression.

**Regression Assumptions**

- independent variable X is fixed (controlled or measured without error)
- dependent variable Y contains error or variability
  - Y is sampled from a normally distributed population
- relationship is linear
- errors in Y:
  - are independent
  - have constant variance \( \sigma^2 \), independent of X
  - have a mean of 0, independent of X
  - are normally distributed

**Mathematical Model**

\[ Y = \alpha + \beta X + \varepsilon \]

For multiple measurements of Y, \( \varepsilon \)'s cancel out, so

\[ \mu_{Y,X} = \alpha + \beta X \] in the population

\( \hat{C} = a + bX \) in the sample

\( Y_i - \hat{C} \) are the errors

\[ b = \frac{SP}{SSX} \]

\[ a = \bar{Y} - b\bar{X} \]
Regression as Orthogonal Contrast

coefficients \( c_i = \frac{X_i - \bar{X}}{\Sigma(X_j - \bar{X})^2} \)

\[ SS = \frac{\left(\sum c_i Y_i\right)^2}{r \Sigma e_i^2} \]

Regression in Replicated Experiments

ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>rt-1</td>
<td>as usual</td>
</tr>
<tr>
<td>Block</td>
<td>r-1</td>
<td></td>
</tr>
<tr>
<td>Treatment</td>
<td>t-1</td>
<td></td>
</tr>
<tr>
<td>Error</td>
<td>(r-1)(t-1)</td>
<td></td>
</tr>
</tbody>
</table>

Subdivide treatment SS

\[ SS_{Regression} = r^2 \cdot SSTrt \]

\[ SS_{Deviation from Regression} = (1-r^2) \cdot SSTrt = SSTrt - SS_{Regr} \]

Use MSE for F-test for both MSRegr and MSDev