# Correlation and Regression

- a *scatterplot* is used to assess the relationship between two variables

- each point shows the values of the two variables $(x_i, y_i)$ measured on the same individual

- look for the overall pattern and for striking deviations from it

- two variables are *associated* if some values of one variable tend to occur more often with some values of the the other variable

- can describe the *form*, *direction* and *strength* of any association

  - form can be *linear* or *nonlinear*, *positive* or *negative*

- sometimes we hope to explain one variable by the other
  - we call them the *response* and *explanatory* variables
  - the response variable is shown on the vertical axis
- we may want to explain or predict the useable volume in board feet/10 of a tree given a measurement at chest height in inches

```
MTB > set c1
DATA> 36 28 28 41 19 32 22 38 25 17 31 20 25 19 39 33 17 37 23 39
DATA> set c2
DATA> 192 113 88 294 28 123 51 252 56 16 141 32 86 21 231 187 22 205 57 265
MTB > name c1 'diameter'
MTB > name c2 'volume'
MTB > plot c2 c1
         -
      300+                                                     *
         -
 volume  -                                               * *
         -                                                 *
         -
      200+                                           * *
         -                                     *
         -
         -                                 *
         -                           *       *
      100+
         -                     *     *
         -               * *   *
         -           *
         -     2   2
        0+
           +---------+---------+---------+---------+---------+------diameter
        15.0      20.0      25.0      30.0      35.0      40.0
```

## Correlation

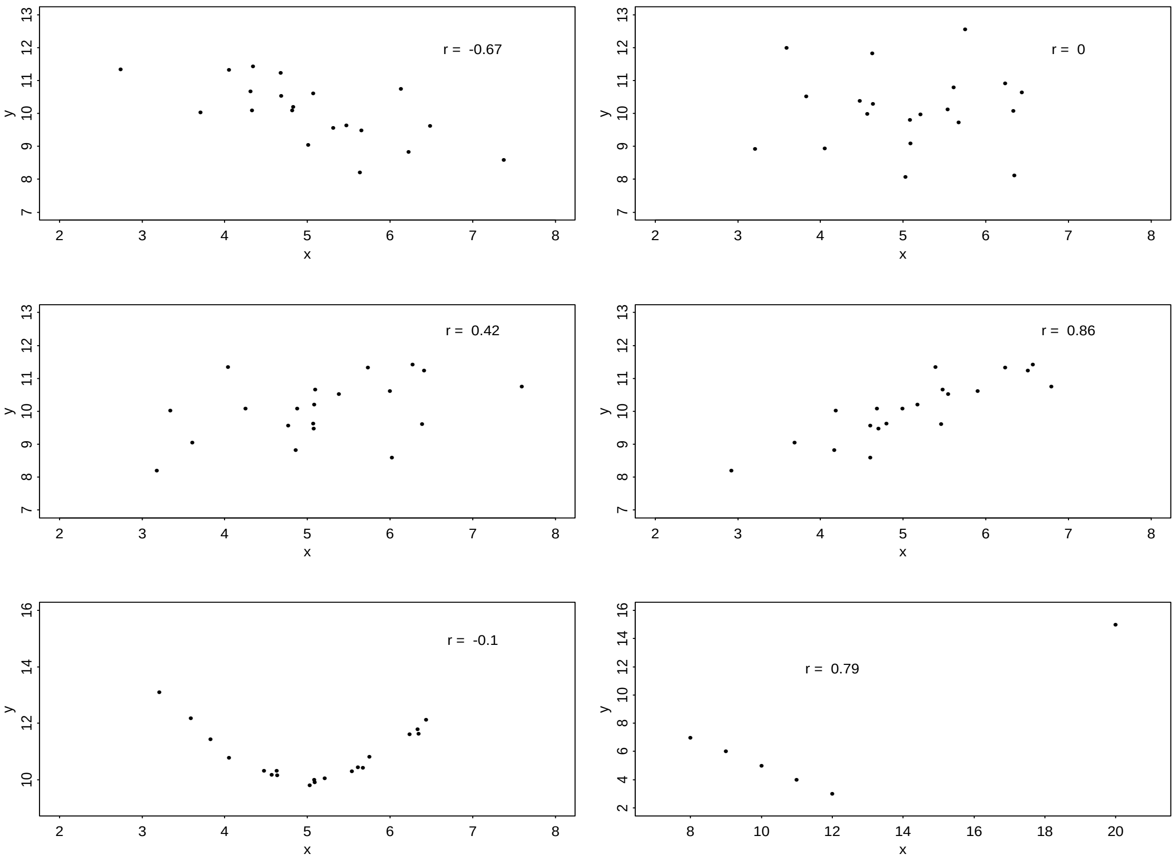
- the *correlation coefficient* measures the direction and strength of the *linear* association between two quantitative variables
- given data $(x_i, y_i), i = 1 \ldots n$, the correlation coefficient is

$$r = \frac{1}{n-1} \sum \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

- the product of the two terms in braces is positive if both $x_i$ and $y_i$ are above or below their means
- $r$ must be between -1 and 1
- $r = 0$ means no linear association
- $r = 1(-1)$ means all points fall on a line with positive (negative) slope
- calculating correlation coefficient in MINITAB

```
MTB > corr c1 c2
 Correlation of diameter and volume = 0.976
```

- some sample plots



- top left - moderately strong negative linear association ($r = -.67$)
- top right - no association ($r = 0$)
- middle left - weak positive association ($r = .42$)
- middle right - strong positive association ($r = .86$)

- bottom left - strong quadratic association (zero linear, $r = 0$)
- bottom right - perfect negative association with one influential outlier ($r = .79$)

Alternative Formulae

- the numerator of the formula for $r$ is
$$SS_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
- so
$$r = \frac{1}{n-1}\frac{SS_{XY}}{s_x s_y}$$
- we can also write
$$SS_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

  and
$$s_x = \sqrt{\frac{SS_{XX}}{n-1}}$$

  so
$$r = \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}}$$

where $SS_{YY}$ is defined similarly to $SS_{XX}$

- note that $SS_{XY}$ can be written in various ways

$$\begin{aligned} SS_{XY} &= \sum_{i=1}^{n}(x_i - \bar{x})y_i \\ &= \sum_{i=1}^{n} x_i(y_i - \bar{y}) \\ &= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \\ &= \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i/n \end{aligned}$$

- the version to use depends on what you are given

Example: To study the effect of ozone pollution on soybean yield, data were collected at four ozone dose levels and the resulting soybean seed yield monitored. Ozone dose levels (in ppm)were reported as the average ozone concentration during the growing season. Soybean yield was reported in grams per plant.

| X<br>Ozone(ppm) | Y<br>Yield (gm/plant) |
|---|---|
| .02 | 242 |
| .07 | 237 |
| .11 | 231 |
| .15 | 201 |

- to calculate the correlation coefficient by hand, we obtain the squares and cross products and their sums

| X | Y | $X^2$ | $Y^2$ | $XY$ |
|---|---|---|---|---|
| .02 | 242 | .0004 | 58564 | 4.84 |
| .07 | 237 | .0049 | 56169 | 16.59 |
| .11 | 231 | .0121 | 53361 | 25.41 |
| .15 | 201 | .0225 | 40401 | 30.15 |

- Column sums: $\sum x_i = .35$, $\sum y_i = 911$, $\sum x_i^2 = .0399$, $\sum y_i^2 = 208,495$, and $\sum x_i y_i = 76.99$

- Means: $\bar{x} = .0875$ and $\bar{y} = 227.95$
- Intermediate terms:

$$
\begin{aligned}
SS_{xx} &= \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - \frac{(\sum x_i)^2}{n} \\
&= .0399 - \frac{(.35)^2}{4} = .009275
\end{aligned}
$$

$$
\begin{aligned}
SS_{yy} &= \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - \frac{(\sum y_i)^2}{n} \\
&= 208,495 - \frac{(911)^2}{4} = 1014.75
\end{aligned}
$$

and

$$
\begin{aligned}
SS_{xy} &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\
&= \sum_i x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} \\
&= 76.99 - \frac{.35(911)}{4} = -2.7225
\end{aligned}
$$

- the correlation coefficient is

$$
\begin{aligned}
r &= \frac{SS_{XY}}{\sqrt{SS_{XX}SS_{YY}}} \\
&= \frac{-2.7225}{\sqrt{.009275(1014.75)}} \\
&= -.8874
\end{aligned}
$$

- there is a strong negative linear association between yield and ozone

Simple Linear Regression

- a line summarizing the relationship between two variables
- has form $y = \beta_0 + \beta_1 x$
  - must choose which variable is the response $y$ and which the explanatory variable $x$
  - $\beta_0$ is the $y$-intercept, the value for $y$ when $x = 0$
  - $\beta_1$ is the slope, the change in $y$ for a unit change in $x$
- can be used to predict value of $y$ for a given $x$
- obtain by minimizing the sum of squares of vertical deviations from the line

$$SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

- note that SSE is a function of $\beta_0$ and $\beta_1$ only because the data $(x_i, y_i)$, $i = 1, \ldots, n$ is known

- the least squares slope has a surprisingly simple formula

$$\hat{\beta}_1 = r\frac{s_y}{s_x}$$

- the fitted intercept is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- the equation of the least squares line is

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x \\ &= \bar{y} + \hat{\beta}_1(x - \bar{x}) \\ &= \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})\end{aligned}$$

- from the latter formula, we see that the fitted value when $x = \bar{x}$ is $\bar{y}$, so the least squares line always goes through the point $\bar{x}, \bar{y}$
- rearranging further, we get

$$\frac{y - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}$$

- this gives another interpretation of the correlation coefficient, namely that it is the slope of the best fitting line if both the $x$ and $y$ variables are standardized

Example: for the tree data, $\bar{y} = 123.0$, $\bar{x} = 28.45$, $r = .976$, $s_y = 91.7$ and $s_x = 8.11$

- the estimated slope is

  $$\hat{\beta}_1 = rs_y/s_x = .976(91.7)/8.11 = 11.036$$

- the estimated intercept is

  $$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} = 123.0 - 11.036(28.45) \\ &= -190.96 \end{aligned}$$

- the fitted line is

  $$volume = -190.96 + 11.036 diameter$$

- if the diameter were 27 inches, we would predict a volume of 107.012 board feet/10)
- these results differ from MINITAB due to round-off error

```
MTB > regress c2 1 c1;
SUBC> residuals c3.
The regression equation is
volume = - 191 + 11.0 diameter

Predictor        Coef       Stdev    t-ratio        p
Constant      -191.12       16.98     -11.25    0.000
diameter      11.0413      0.5752      19.19    0.000

s = 20.33       R-sq = 95.3%     R-sq(adj) = 95.1%
```

```
Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     1      152259      152259    368.43    0.000
Error         18        7439         413
Total         19      159698
```

Example: For the ozone data,

- $\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = -293.531$
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 227.95 - (-293.531)(.0875) = 253.434$
- the least squares line is

$$\widehat{yield} = 253.434 - 293.531 ozone$$

Derivation of formulae for intercept and slope

- those who have taken calculus will know that one can use derivatives to find the maximum or minimum of a function
- in this case there are two variables $\beta_0$ and $\beta_1$, and so both *partial* derivatives can be set to zero and solved
- the partial derivatives are

  $$\frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)$$

  and

  $$\frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i)$$

- when equated to zero and rearranged, these give the so-called "normal equations".

  $$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

and

$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

- (the term 'normal' here has nothing to do with the normal distribution, but rather to the geometric idea of orthogonality or perpendicularity)
- the two normal equations are solved simultaneously to obtain

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- there is a second derivation which doesn't require calculus
- notice that SSE is a quadratic function of $\beta_0$ and $\beta_1$
- first consider $\beta_1$ to be fixed and find the value of $\beta_0$ which minimizes SSE

- completing the square and summing gives

  $$SSE = A\beta_0^2 + B\beta_0 + C$$

  where

  $$\begin{aligned} A &= n \\ B &= 2\beta_1 \sum x_i - 2 \sum y_i \\ C &= \sum y_i^2 - 2\beta_1 \sum x_i y_i + \beta_1^2 \sum x_i^2 \end{aligned}$$

- the minimum of a quadratic occurs at $-B/2A$, so whatever the value of $\beta_1$ the best choice for $\beta_0$ is

  $$\begin{aligned} \beta_0 &= \frac{-2\beta_1 \sum x_i + 2 \sum y_i}{2n} \\ &= \bar{y} - \beta_1 \bar{x} \end{aligned}$$

- now substitute this choice into SSE, so that it is now a quadratic function of $\beta_1$ only

  $$SSE = \sum_{i=1}^{n} (y_i - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_i)^2$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}(y_i - \bar{y} - \beta_1(x_i - \bar{x}))^2 \\
&= A\beta_1^2 + B\beta_1 + C
\end{aligned}
$$

- where now

$$
\begin{aligned}
A &= SS_{XX} \\
B &= -2SS_{XY} \\
C &= SS_{YY}
\end{aligned}
$$

- the minimum occurs at

$$\hat{\beta}_1 = \frac{-B}{2A} = \frac{SS_{XY}}{SS_{XX}}$$

- substituting in $\beta_0$ gives

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_1\bar{x}$$