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 cor-

relation 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

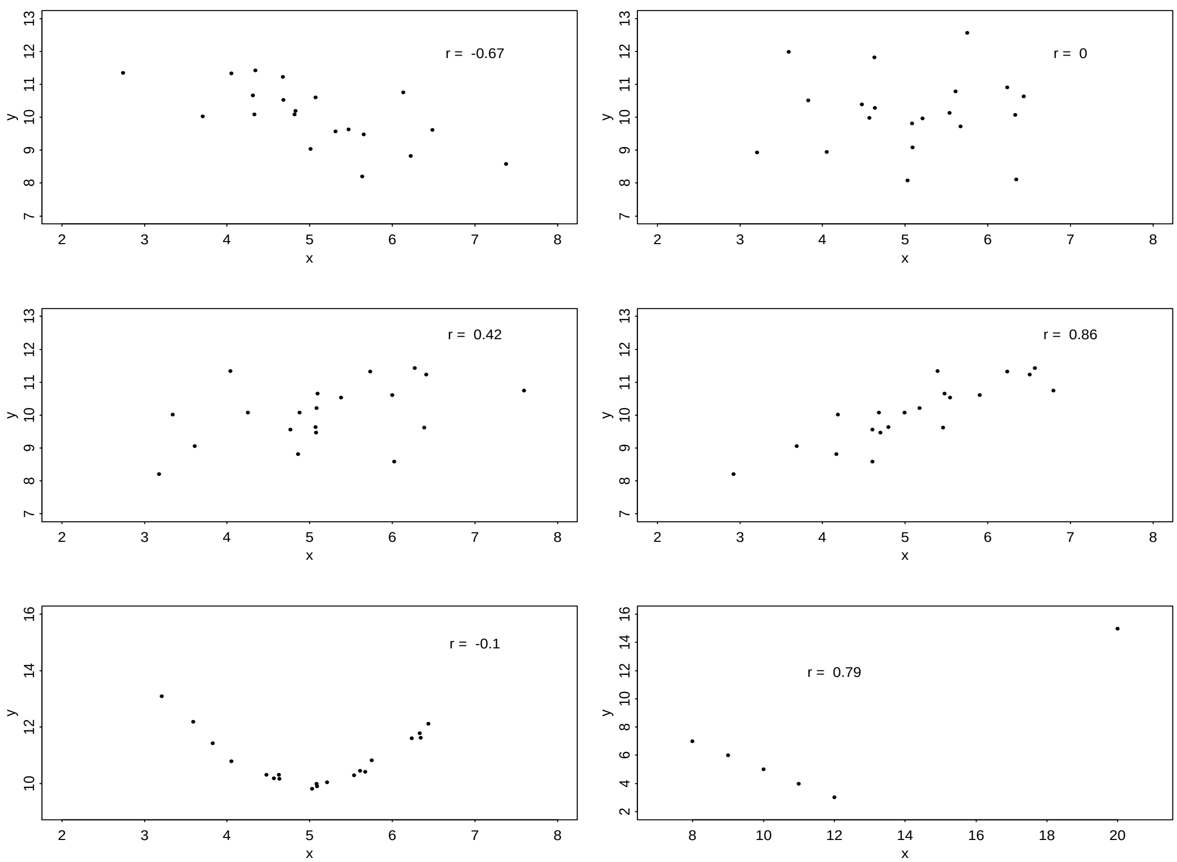
- $r = 0$ means no linear association

- $r = 1(-1)$ 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 ($r = .79$)

Least-Squares Regression

- a line summarizing the relationship between two variables
- has form $y = \beta_0 + \beta_1 x$
  - must choose response y and explanatory variable x
  - β_0 is the y-intercept
  - β_1 is the slope
- can be used to predict value of y for a given x
- fit to data by minimizing the sum of squares of vertical deviations from the line
$$\sum (y_i - \beta_0 - \beta_1 x_i)^2$$
- fitted slope
$$\hat{\beta}_1 = r \frac{s_y}{s_x}$$

- fitted intercept

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

- 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

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

- the estimated intercept is

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

- 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
```

- the fitted line always passes through $(\bar{x}, \bar{y})$
- r^2 measures the proportion of the variation in y which has been explained by the regression on x
- the initial variation is $\sum (y_i - \bar{y})^2$ (TSS $=$ total sum of squares)

- the final variation is $\sum (y_i - \hat{y}_i)^2$ (SSE $=$ sum of squares of the errors)
- the amount explained is the difference - we can write this $\sum (\hat{y}_i - \bar{y})^2$ (SSR $=$ regression sum of squares)
- the proportion explained is

$$r^2 = \frac{SSR}{TSS}$$

- the residuals $y_i - \hat{y}_i$ add to zero and should be randomly scattered when plotted against x_i

```
MTB > plot c3 c1

         -
         -                                                     *
       25+     *                                         * *
         -     *
 C3      -                                     *
         -         *
         -
        0+         * *   *     *
         -                 *         *
         -                                 *           *   *
         -                                           *
         -
      -25+
         -                     *     *
         -
         -                                   *
         -
           +---------+---------+---------+---------+---------+------C1
        15.0      20.0      25.0      30.0      35.0      40.0
```

- there is clearly some curvature here
- one remedy is to add a quadratic term in the equation, giving

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

- MINITAB can fit this too

```
MTB > let c3 = c1**2


MTB > regress c2 2 c1 c3;
SUBC> residuals c4.

The regression equation is
volume = 29.7 - 5.62 diameter + 0.290 C3

Predictor       Coef       Stdev    t-ratio        p
Constant       29.74       51.39       0.58    0.570
diameter      -5.620       3.792      -1.48    0.157
C3           0.29037     0.06572       4.42    0.000

s = 14.27       R-sq = 97.8%     R-sq(adj) = 97.6%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    2      156236       78118    383.54    0.000
Error        17        3463         204
Total        19      159698

SOURCE       DF      SEQ SS
diameter      1      152259
C3            1        3976
```

```
MTB > plot c4 c1

 C4      -
         -                                     *
         -
       20+
         -                     *                         *
         -                           *                     *
         -                                 *                   *
         -     *         * *
        0+         * *
         -     *
         -         *
         -                           *               *
         -                     *                       *
      -20+                                                 *
         -                                   *
         -
         -
           +---------+---------+---------+---------+---------+------diameter
        15.0      20.0      25.0      30.0      35.0      40.0
```

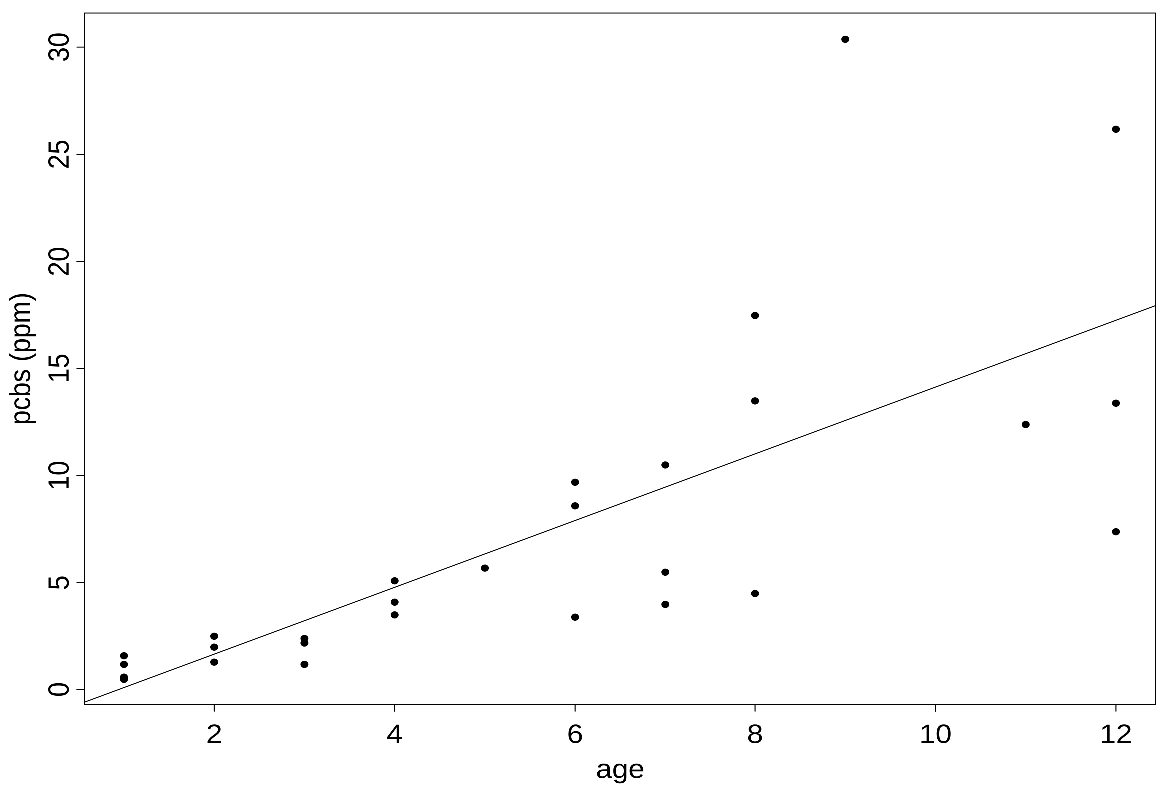
- the new residual plot shows no curvature
- but there is a tendency for the residuals to be larger at larger diameters
- this is harder to fix!

Comments

- correlation measures only linear association - straight line fits only make sense when the data is linear - PLOT!

- extrapolation - using a model outside the range of the data - is dangerous - the form of the relationship may change
- correlation and regression are not resistant - see bottom right panel is earlier plot
- *lurking* variables may make a correlation or regression misleading

- nonlinear transformations on either the response or the explanatory variable or both can simplify the form of the association
- consider the PCB concentration in Cayuga Lake Trout, plotted against the age of the fish

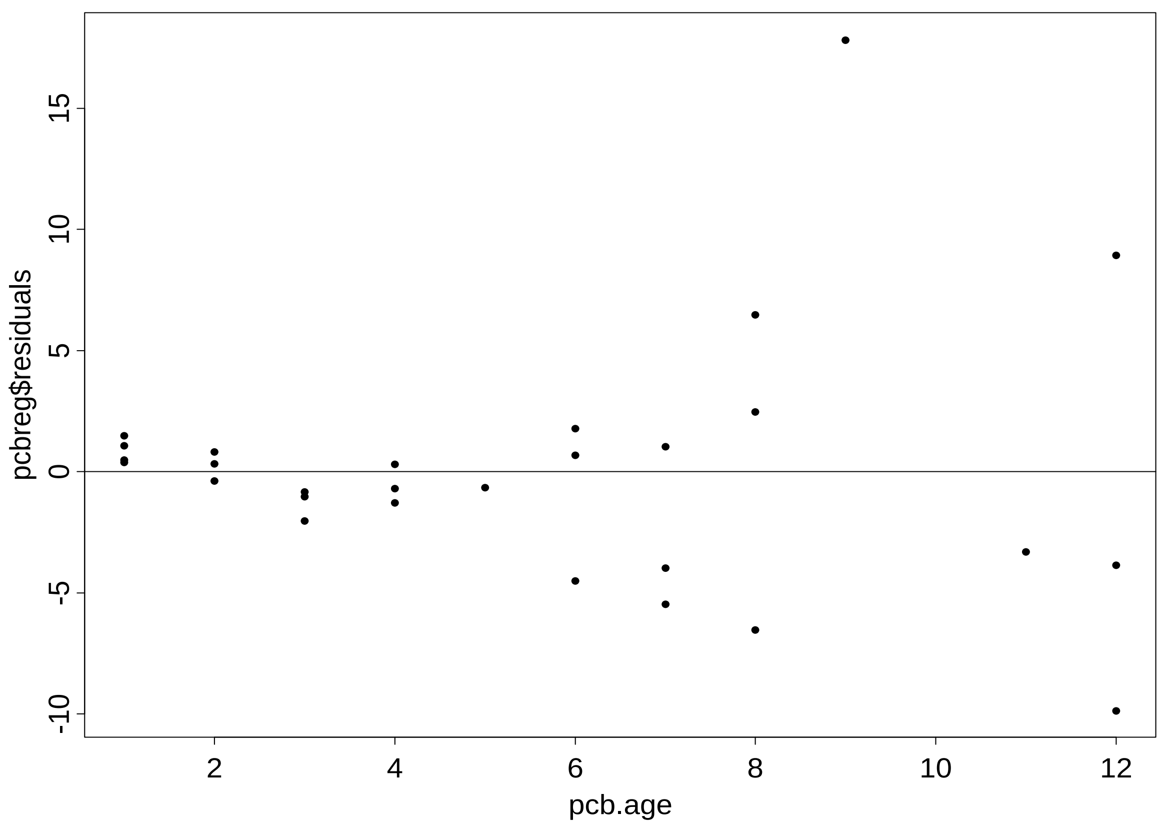


- the fitted least squares line is

$$PCB = -1.45 + 1.56age$$

with $R^2 = .54$

- the residuals, however show problems

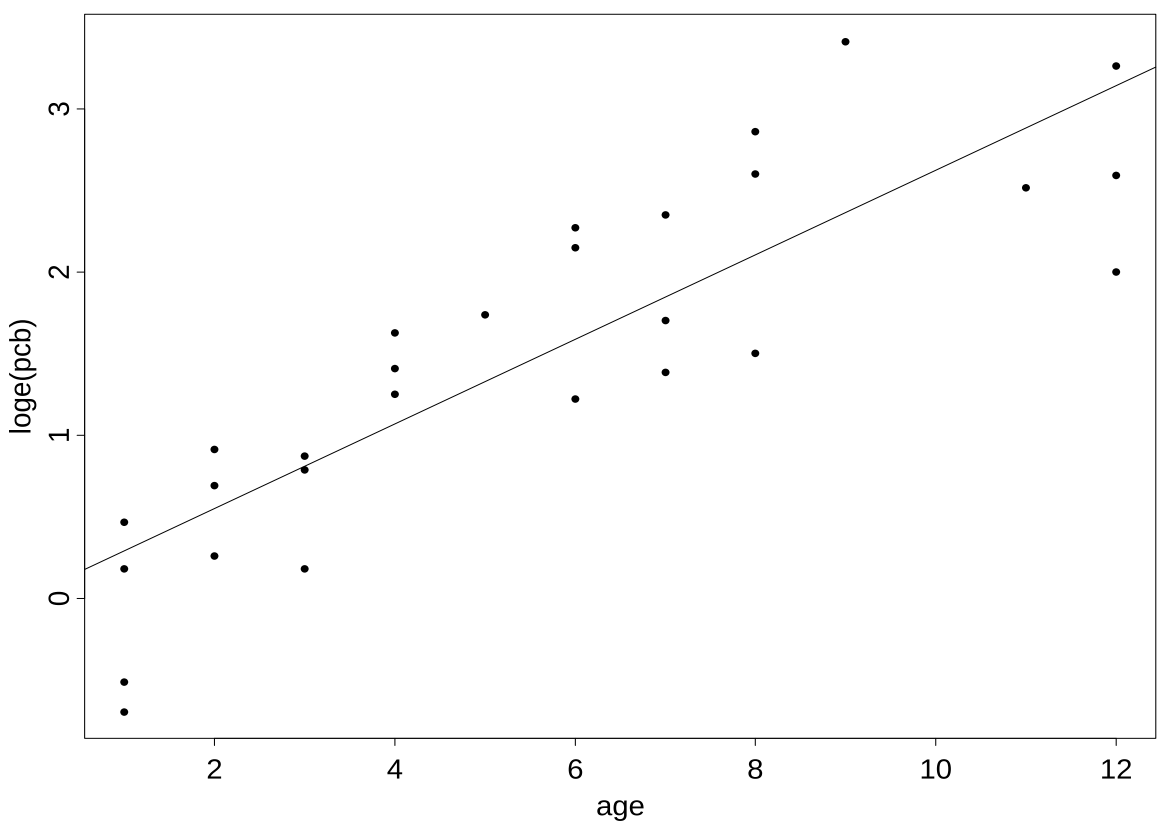


- the residuals are larger at larger ages

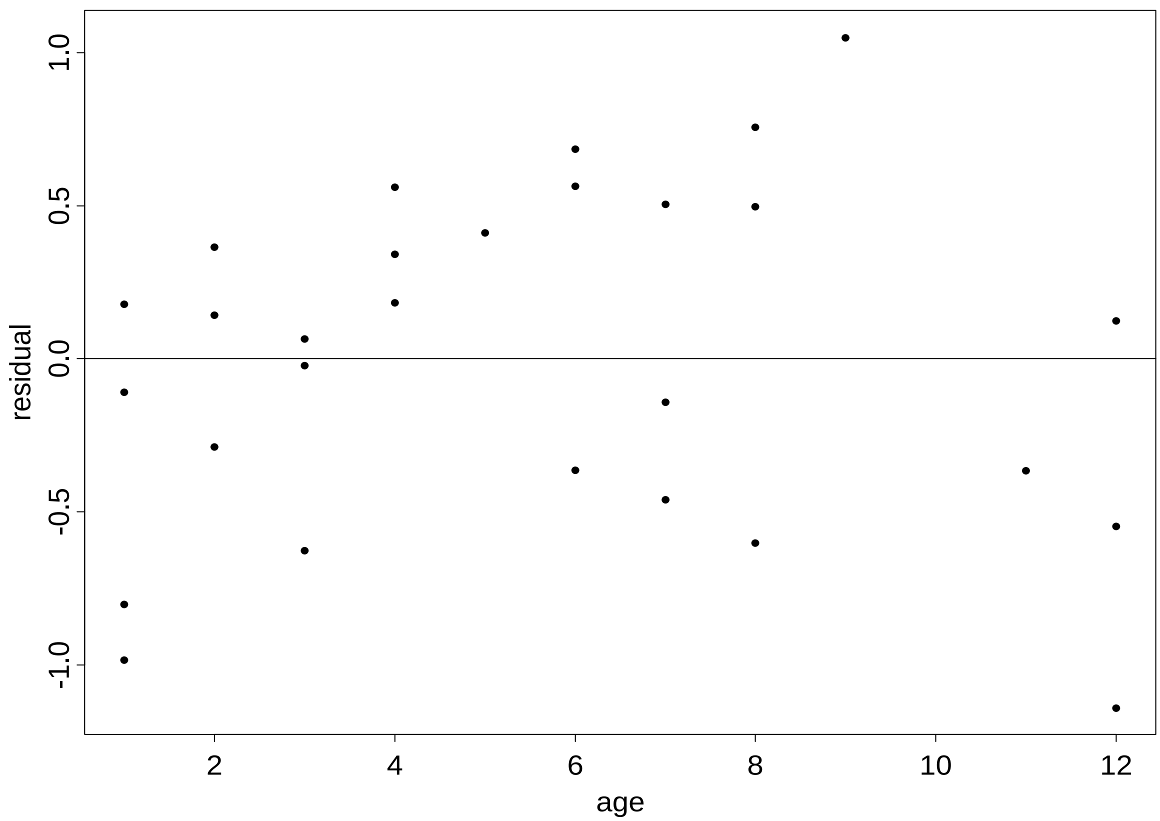
- there is some curvature in the plot
- the plot of log(PCB) versus age, with least squares line is shown
- the least squares fit is

$$log(PCB) = .03 + .259age$$

with $R^2 = .73$



- the residual plot shows even spread for all ages



- there is a very slight suggestion of curvature

- the model says

$$PCB = e^{.03+.259age}$$

- comparing model predictions at age and
  $age\ +1$ gives

$$\frac{PCB_{age+1}}{PCB_{age}} = \frac{e^{.03+.259(age+1)}}{e^{.03+.259age}} = e^{.259} = 1.3$$

so

$$PCB_{age+1} = 1.3PCB_{age}$$

- this is an example of **exponential growth**
  - where growth increases by a fixed percentage of the previous total
  - linear growth increases by a fixed amount
  - growth of bacteria, compound interest are both examples of exponential growth
- general forms

$$y = ab^t$$

or

$$y = ae^{bt}$$

- made linear by logarithmic transformation

$$log(y) = log(a) + tlog(b)$$

or

$$log(y) = log(a) + bt$$