Properties of the least squares fit

1. the fitted line passes through $(\bar{x}, \bar{y})$

   - see this by substituting $x_i = \bar{x}$ into the fitted line

   $$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x})$$

2. the mean of the fitted values is the same as the mean of the observed responses

   - the mean of the fitted values is

   $$\begin{aligned} \bar{\hat{y}} &= \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i \\ &= \frac{1}{n}\sum_{i=1}^{n}(\bar{y} + \hat{\beta}_1(x_i - \bar{x})) \\ &= \bar{y} + \frac{\hat{\beta}_1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \bar{y} \end{aligned}$$

3. the mean of the residuals is zero

   - the residuals are $\hat{e}_i = y_i - \hat{y}_i$
   - so, using (2) above

$$\begin{aligned} \bar{\hat{e}} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \\ &= \bar{y} - \bar{\hat{y}} = 0 \end{aligned}$$

4. the residuals have zero correlation with the predictor

   - we can show $SS_{\hat{e}X} = 0$

$$\begin{aligned} SS_{\hat{e}X} &= \sum_{i=1}^{n}(\hat{e}_i - \bar{e})(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))(x_i - \bar{x}) \end{aligned}$$

$$= \quad SS_{XY} - \hat{\beta}_1 SS_{XX} = 0$$

- the residuals have zero correlation with the fitted values
- we can show $SS_{\hat{e}\hat{y}} = 0$

$$\begin{aligned}
SS_{\hat{e}\hat{y}} &= \sum_{i=1}^{n} \hat{e}_i(\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^{n} (y_i - \hat{y}_i)\hat{\beta}_1(x_i - \bar{x}) \\
&= \sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))\hat{\beta}_1(x_i - \bar{x}) \\
&= \hat{\beta}_1 SS_{XY} - \hat{\beta}_1^2 SS_{XX} \\
&= \hat{\beta}_1 SS_{XY} - \hat{\beta}_1 SS_{XY} = 0
\end{aligned}$$

Ozone example: the fitted values, residuals, sums and crossproducts are shown below

	x_i	y_i	$\hat{y}_i$	$\hat{e}_i = y_i - \hat{y}_i$	$\hat{e}_i x_i$
	.02	242	247.563	-5.563	-.1113
	.07	237	232.887	4.113	.28791
	.11	231	221.146	9.854	1.0840
	.15	201	209.404	-8.404	-1.2606
Sum		911	911	0	0

- the observed and fitted responses have the same sum
- the residuals have zero sum
- the correlation between residuals and predictors will be zero because the sum of cross products is zero

Plotting residuals to assess fit

- from (3) above, the residuals have zero mean, and from (4) and (5) they are uncorrelated with the predictor x and the fitted values $\hat{y}$
- a scatterplot of the residuals versus x should show random scatter about 0, with no linear association with x
- the scatterplot of residuals versus fitted values should be similar
- various problems can be revealed from the plot of $\hat{e}$ versus x or $\hat{y}$
  - curvature indicates that the form of the model is not correct
    - this can be fixed by adding the term x^2 to the model or by transforming the response variable

- the magnitude of the residuals may increase or decrease with the predictor - sometimes called ‘fanning’ out
  - when we use least squares and minimize SSE, we give equal weight to all n deviations
  - this implicitly assumes that the deviations are all roughly the same size
  - this problem can be fixed using a weighted least squares criterion (giving smaller weight to the larger deviations) or by transformation

Example: Lumber example - useable volume versus diameter at chest height

```
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

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

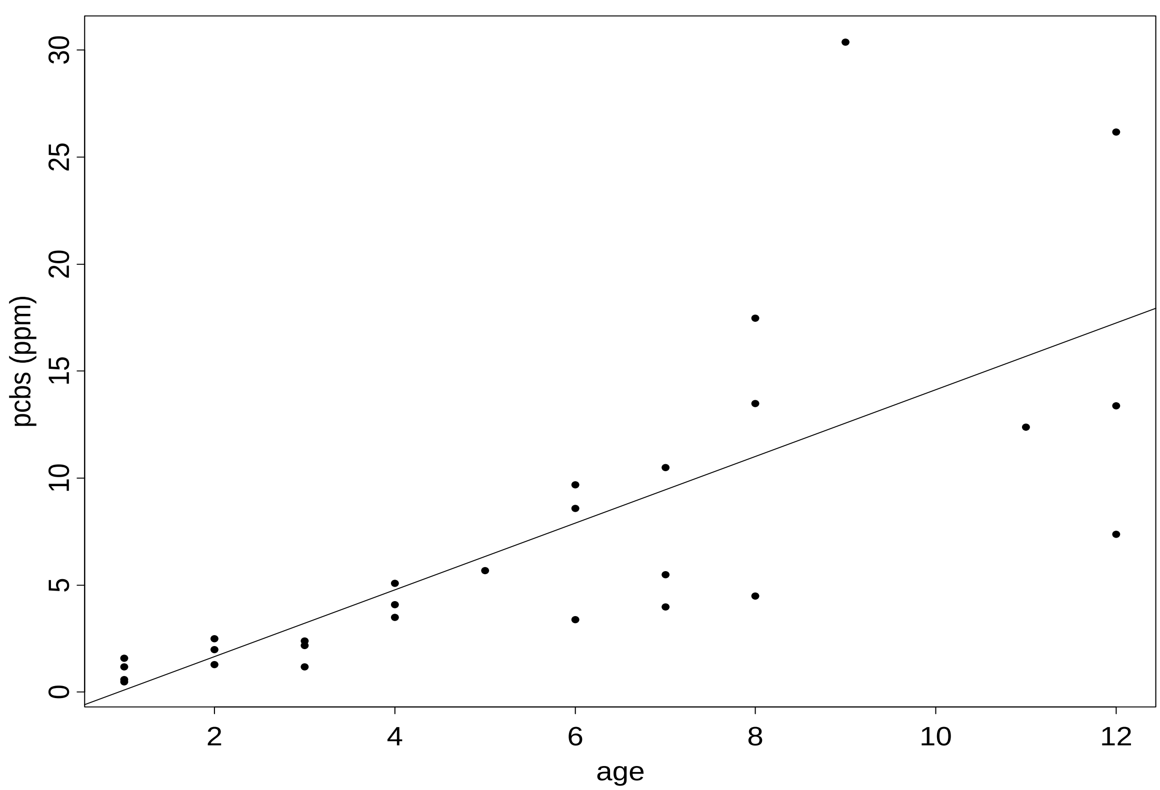
- MINITAB can fit this too

```
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

Example: PCBs in lake trout

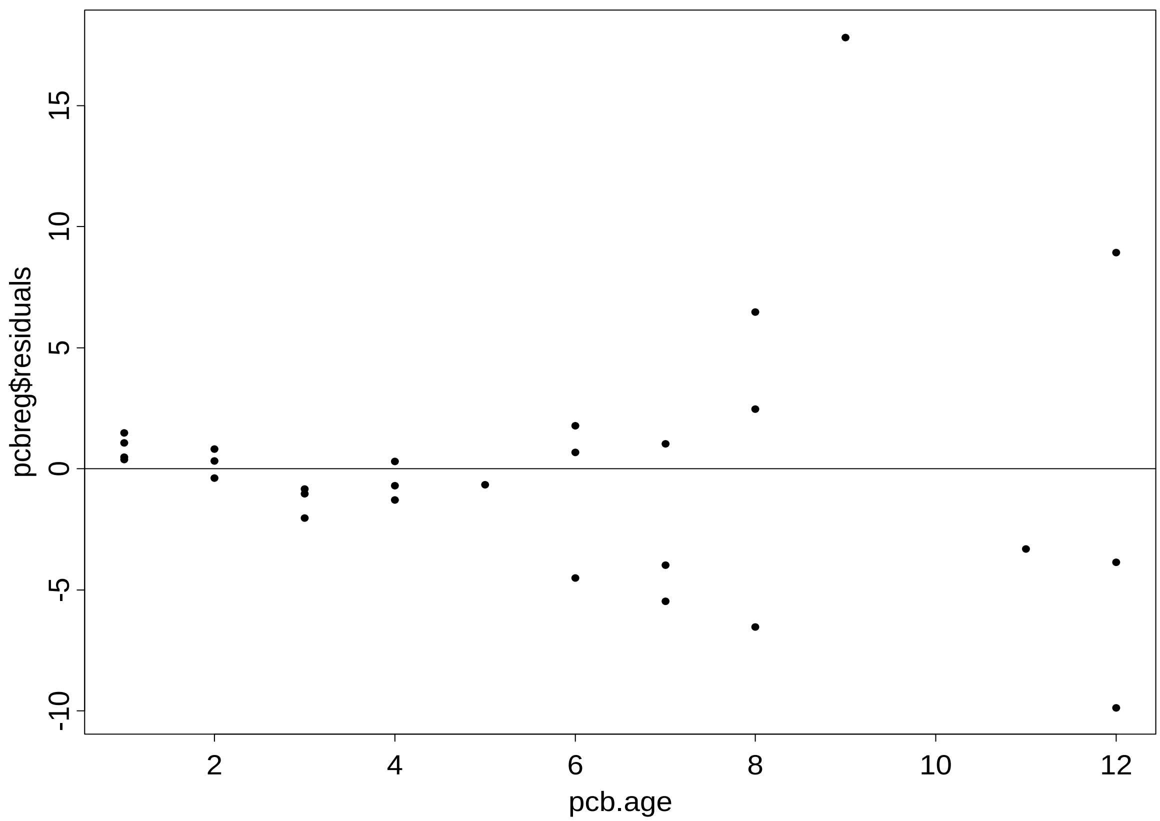
- consider the PCB concentration in Cayuga Lake Trout, plotted against the age of the fish



- the fitted least squares line is

$$PCB = -1.45 + 1.56 age$$

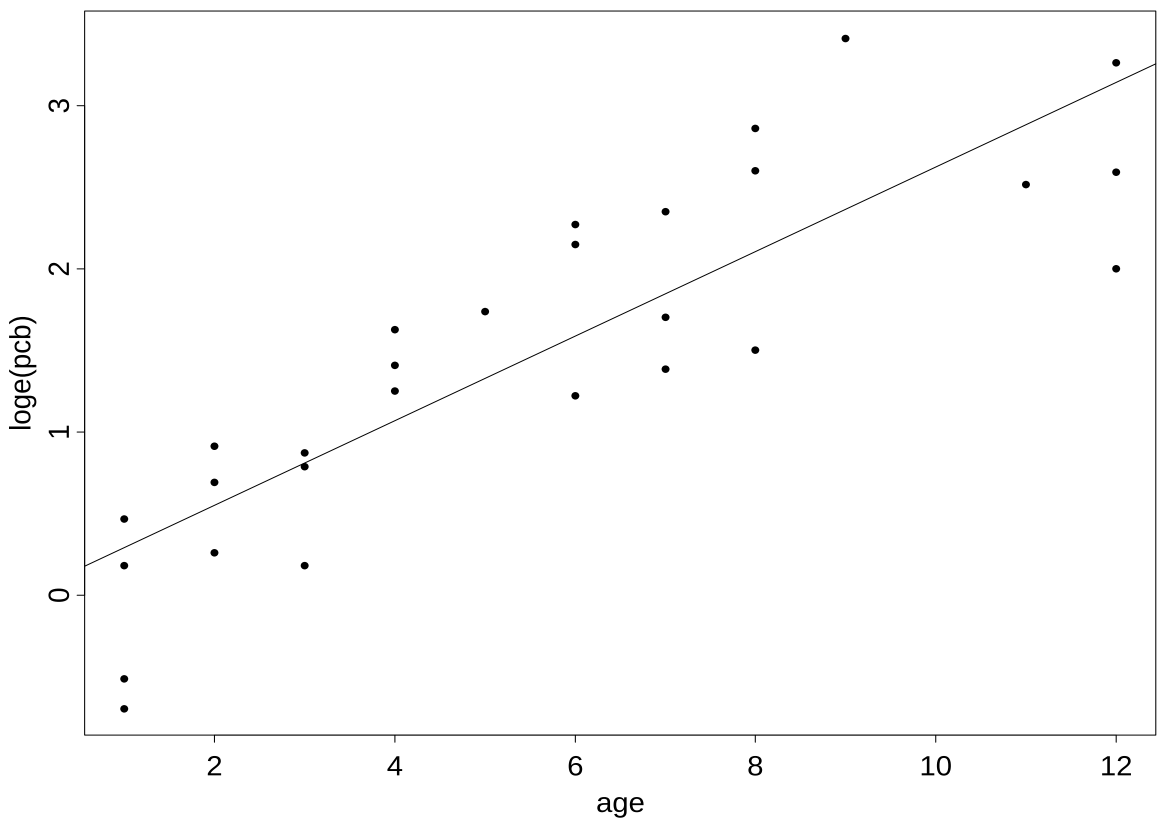
- the residuals, however show problems



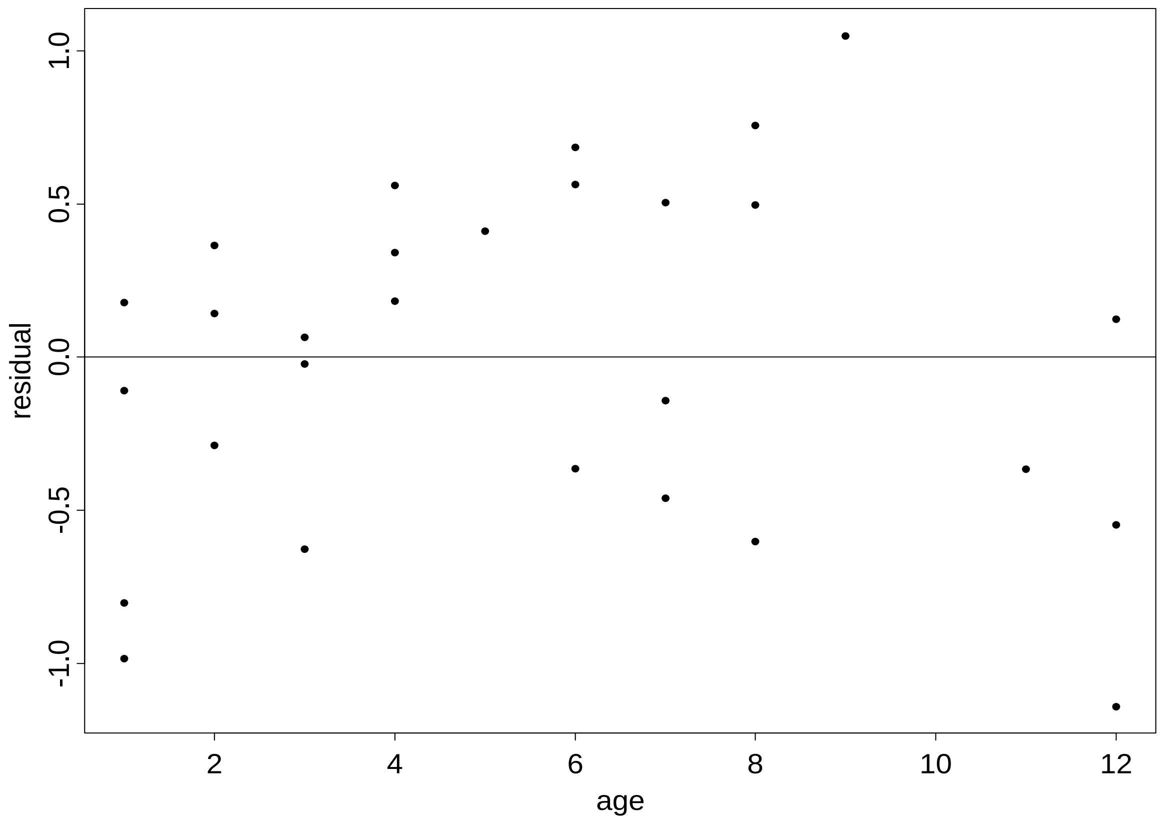
- the residuals are larger at larger ages
- there is some curvature in the plot
- the plot of log(PCB) versus age, with least squares line is shown

- the least squares fit is

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



- the residual plot shows even spread for all ages



- 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