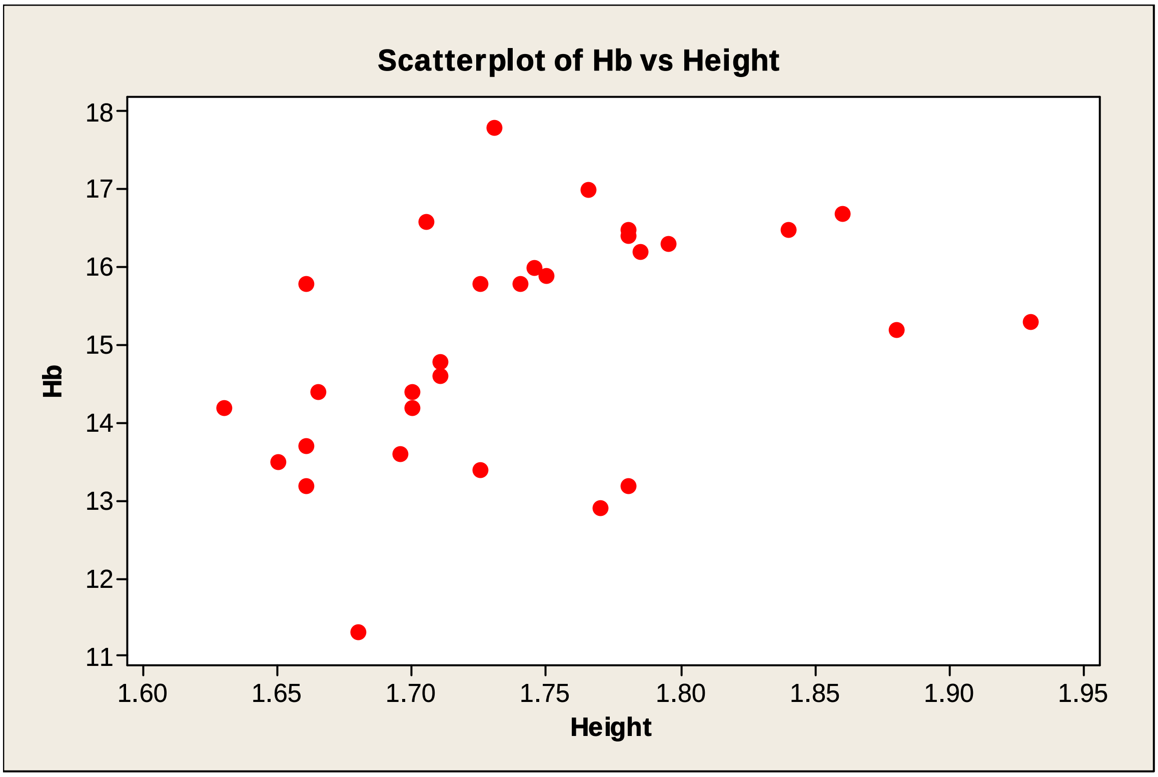
MRes in Medical Statistics
MMB8028

School of Mathematics and Statistics

Practical session on regression and correlation: outline solutions

1.

Select the Scatterplot… item on the **Graph** menu. Then choose Simple and make Hb the y-variable and Height the x-variable: clicking on OK gives the following:



Using the Regression… sub-item under the Regression item in the **Stat** menu with Hb as the Response and Height as the Predictor the following output is obtained:

Regression Analysis: Hb versus Height

```
The regression equation is
Hb = - 0.70 + 9.05 Height


Predictor     Coef  SE Coef      T      P
Constant    -0.704    6.250  -0.11  0.911
Height       9.048    3.589   2.52  0.018


S = 1.38530   R-Sq = 18.5%   R-Sq(adj) = 15.6%
```

(The remainder of output is not needed and can be ignored)

The equation suggests that the mean Hb increases by 9.05 g/dl for each 1m increase in height (of course, in any report it would be preferable to say mean Hb increases by 0.905 g/dl for each 0.1m increase in height, or something similar). The P-value of 0.018 accompanying `Height` shows that the increasing trend of Hb with height observed in the plot cannot be ascribed to chance alone. The spread about the line has SD 1.385 g/dl and the correlation is $\sqrt{0.185}$ = 0.43. The `R-sq(adj)` can be ignored as its main value is in multiple regression.

[n.b. in this regression, and most others, the P-value ascribed to `Constant` (i.e. the estimate of the intercept, is of no interest and can be ignored)]

However, should Hb be dependent on height? This seems unlikely.

If separate columns for females (called Hbf and Htf) and for males (Hbm and Htm) are created then the regression can easily be repeated separately for males and females.

Performing the regression on the females gives the following (abridged as above):

Regression Analysis: Hbf versus Htf

```
The regression equation is
Hbf = 23.0 - 5.41 Htf


Predictor    Coef  SE Coef      T      P
Constant    22.98    11.07   2.08  0.058
Htf        -5.412    6.536  -0.83  0.423


S = 1.03842   R-Sq = 5.0%   R-Sq(adj) = 0.0%
```

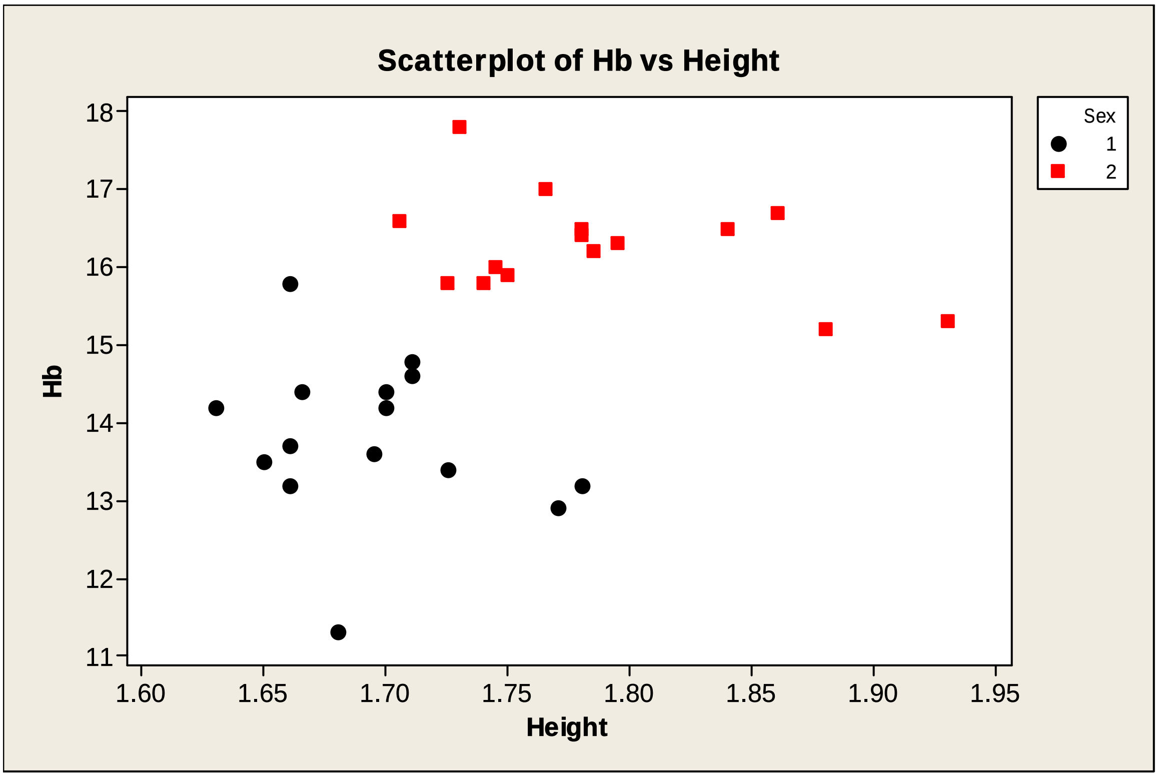
and the corresponding analysis on males is:

Regression Analysis: Hbm versus Htm

```
The regression equation is
Hbm = 24.1 - 4.40 Htm


Predictor    Coef  SE Coef      T      P
Constant   24.126    4.611   5.23  0.000
Htm        -4.397    2.578  -1.71  0.112


S = 0.615839   R-Sq = 18.3%   R-Sq(adj) = 12.0%
```

In both analyses the P-value for the slope is not significant and means that the observed trends of Hb with height within each sex can be ascribed to chance. The P-value attached to the `Constant` term is of no interest.

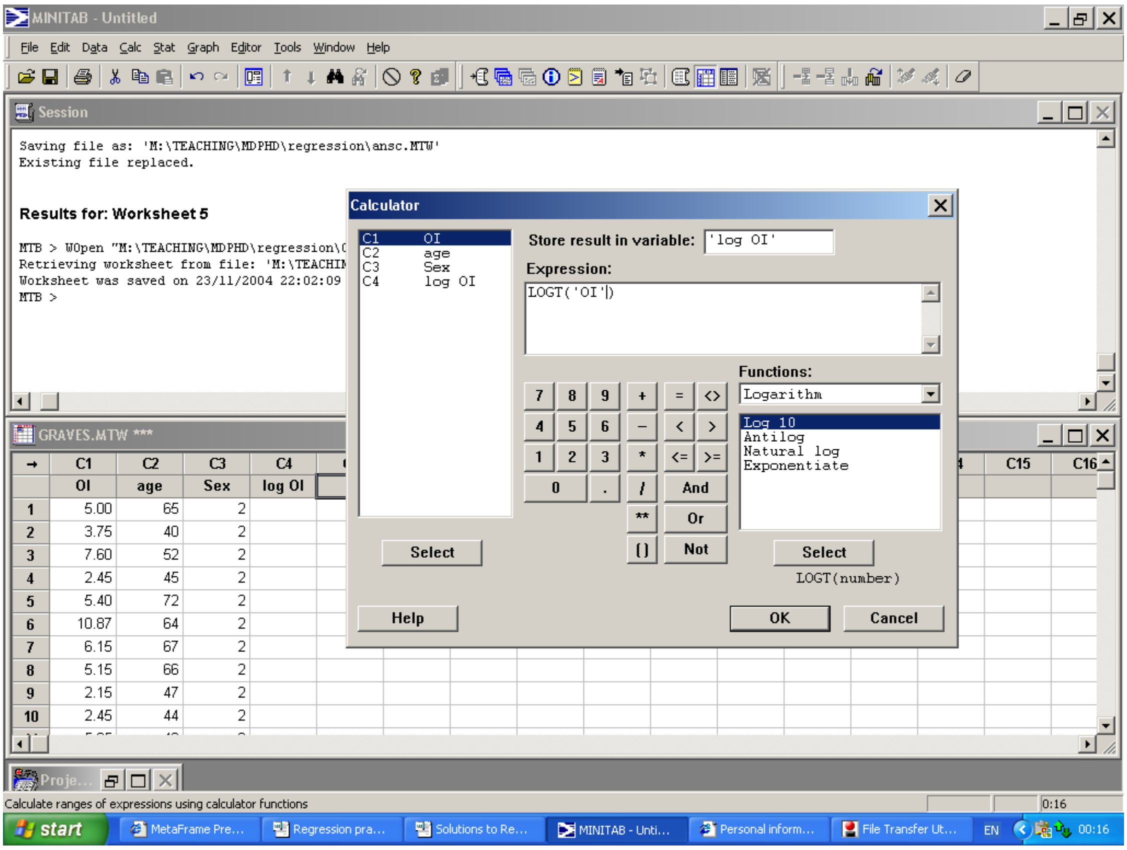
These analyses are all very well but the fundamental rule of regression analysis is always to plot the data. If this is done, but now distinguishing males and females on the plot by using different symbols. You can do this by the same sequence of commands as above but, rather than selecting Simple, select With Groups. Make sure you enter Sex in the Categorical variables for grouping (0-3): box., we obtain the following:



This plot shows the same data as before but with separate symbols for males (🟥) and females (●). The original result was due to the fact that males tend to be taller and have higher haemoglobin concentrations, not that tall people *per se* have higher Hbs. In general, just because there appears to be a relation between Y and X does not mean that the observed relation is useful or meaningful. Once a third (or more) variable, Z, is taken into account the relation between Y and X may disappear. It is also possible for apparently unrelated variables to "become" related when a further variable is taken into account.

2.

The log (to base 10) of OI can be calculated and stored in a column (which has been labelled `log OI`) using the Calculator… item on the Calc menu. To do this, the dialogue box should look as below before you click on OK.



MINITAB - Untitled

Session

```
Saving file as: 'M:\TEACHING\MDPHD\regression\ansc.MTW'
Existing file replaced.
```

Results for: Worksheet 5

```
MTB > WOpen "M:\TEACHING\MDPHD\regression\(
Retrieving worksheet from file: 'M:\TEACHIN
Worksheet was saved on 23/11/2004 22:02:09
MTB >
```

Calculator

```
C1      OI
C2      age
C3      Sex
C4      log OI
```

Store result in variable: 'log OI'

Expression: LOGT('OI')

Functions: Logarithm

Log 10
Antilog
Natural log
Exponentiate

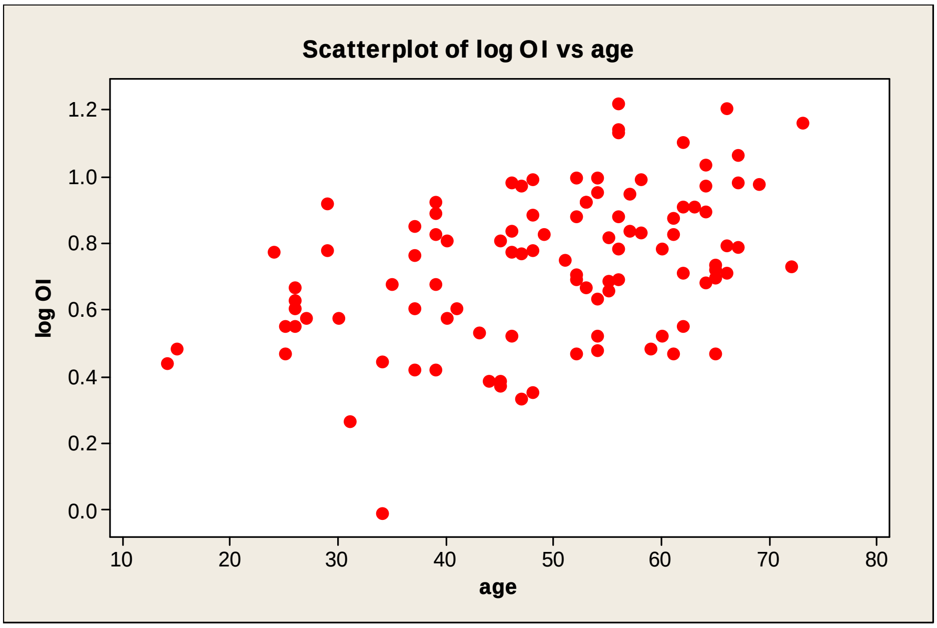
LOGT(number)

GRAVES.MTW ***

→	C1	C2	C3	C4
	OI	age	Sex	log OI
1	5.00	65	2	
2	3.75	40	2	
3	7.60	52	2	
4	2.45	45	2	
5	5.40	72	2	
6	10.87	64	2	
7	6.15	67	2	
8	5.15	66	2	
9	2.15	47	2	
10	2.45	44	2	

Calculate ranges of expressions using calculator functions

{A quicker way is to enter the command `Let C4 = logt(OI)` at the Minitab prompt (MTB>) in the Session window.}



Scatterplot of $\log_{10}$ OI against age.

The scatterplot of log OI against age (shown above) indicates that there is a tendency for the log ophthamic index to increase with age. As the OI is a measure of deteriorating visual performance (larger OI → poorer sight) this is probably not surprising. The graph is obtained in the same way as you obtained the first plot in question 1.

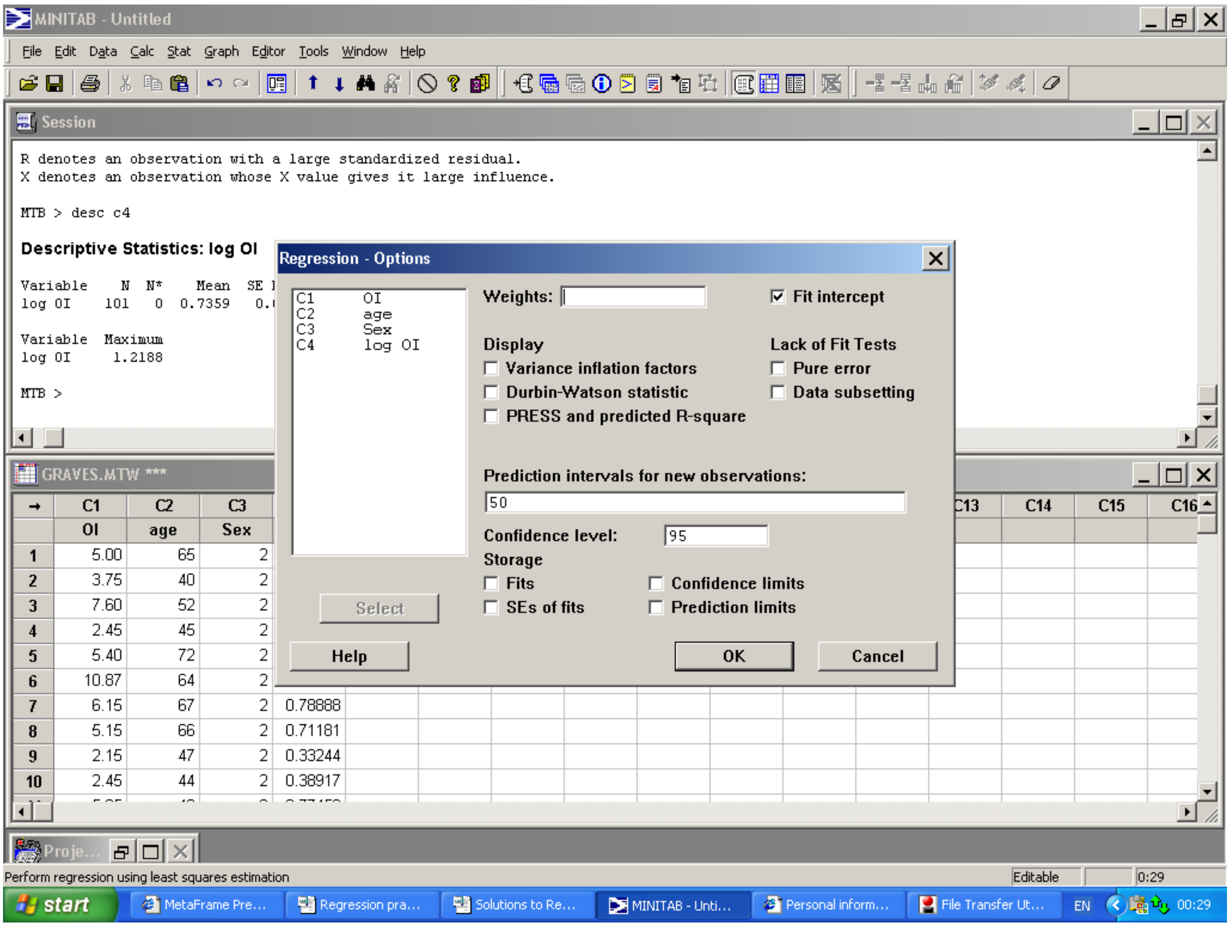
Regressing log OI on age gives the following.

Regression Analysis: log OI versus age

```
The regression equation is
log OI = 0.369 + 0.00743 age


Predictor       Coef   SE Coef     T      P
Constant     0.36945   0.07767  4.76  0.000
age         0.007431  0.001520  4.89  0.000


S = 0.204855   R-Sq = 19.5%   R-Sq(adj) = 18.6%
```

The P-value for age, which to 3 d.p. is 0.000, reveals clearly that the trend of increasing mean log OI with age is not due to chance. The mean log OI increases by 0.00743 for each age increase of 1 year and is distributed about this mean with SD 0.205 (if we calculate the SD of log OI, without taking age into account we obtain a value of 0.227).

The predicted log OI at age 50 is found from the regression equation as 0.369+0.00743×50 = 0.741. However, computing the limits of log OI for a 50-year-old by hand is not so straightforward and the easiest way is to redo the regression calculation, with some further options selected. Repeat the commands you entered above until you reach the Regression dialogue box. Now click on Options… and enter 50 in the Prediction intervals for new observations box, as shown below.



```
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

MTB > desc c4

Descriptive Statistics: log OI

Variable    N  N*    Mean  SE I
log OI    101   0  0.7359   0.(

Variable  Maximum
log OI     1.2188

MTB >
```

Regression - Options

```
C1    OI
C2    age
C3    Sex
C4    log OI
```

Weights:

☑ Fit intercept

Display

□ Variance inflation factors

□ Durbin-Watson statistic

□ PRESS and predicted R-square

Lack of Fit Tests

□ Pure error

□ Data subsetting

Prediction intervals for new observations: 50

Confidence level: 95

Storage

□ Fits

□ SEs of fits

□ Confidence limits

□ Prediction limits

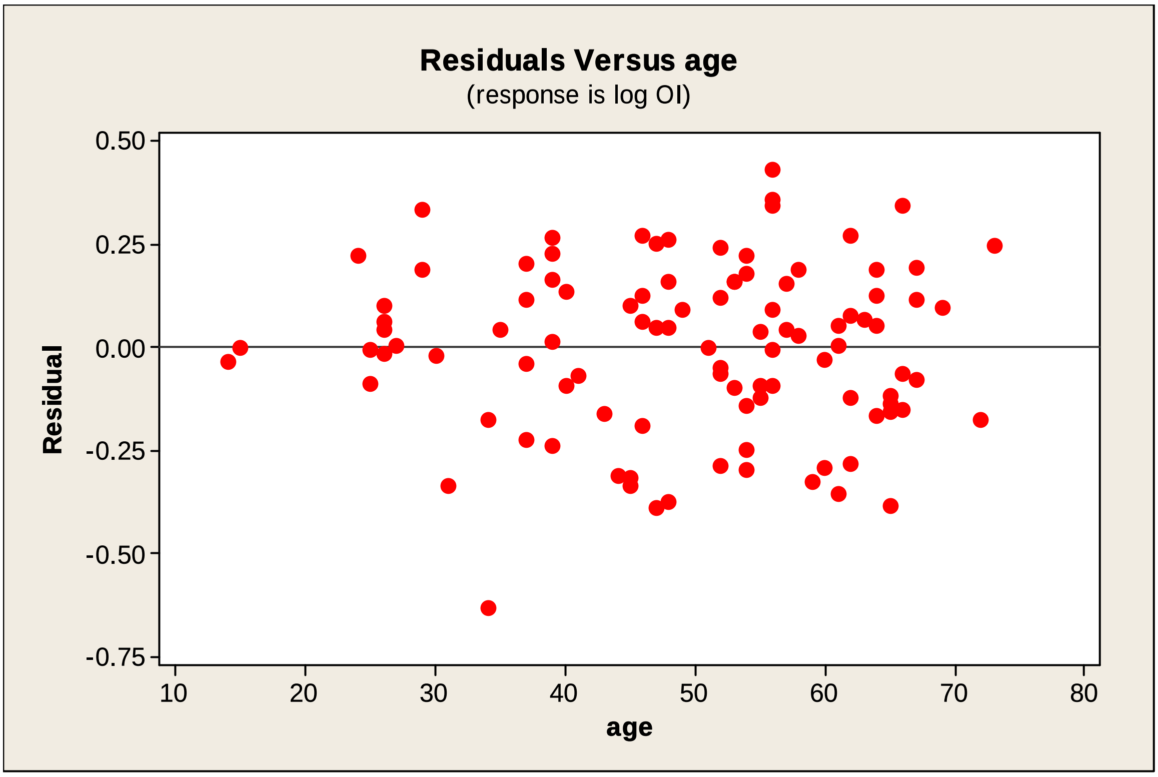
→	C1	C2	C3	
	OI	age	Sex	
1	5.00	65	2	
2	3.75	40	2	
3	7.60	52	2	
4	2.45	45	2	
5	5.40	72	2	
6	10.87	64	2	
7	6.15	67	2	0.78888
8	5.15	66	2	0.71181
9	2.15	47	2	0.33244
10	2.45	44	2	0.38917

This gives the usual regression output with the following new lines towards the end of the output:

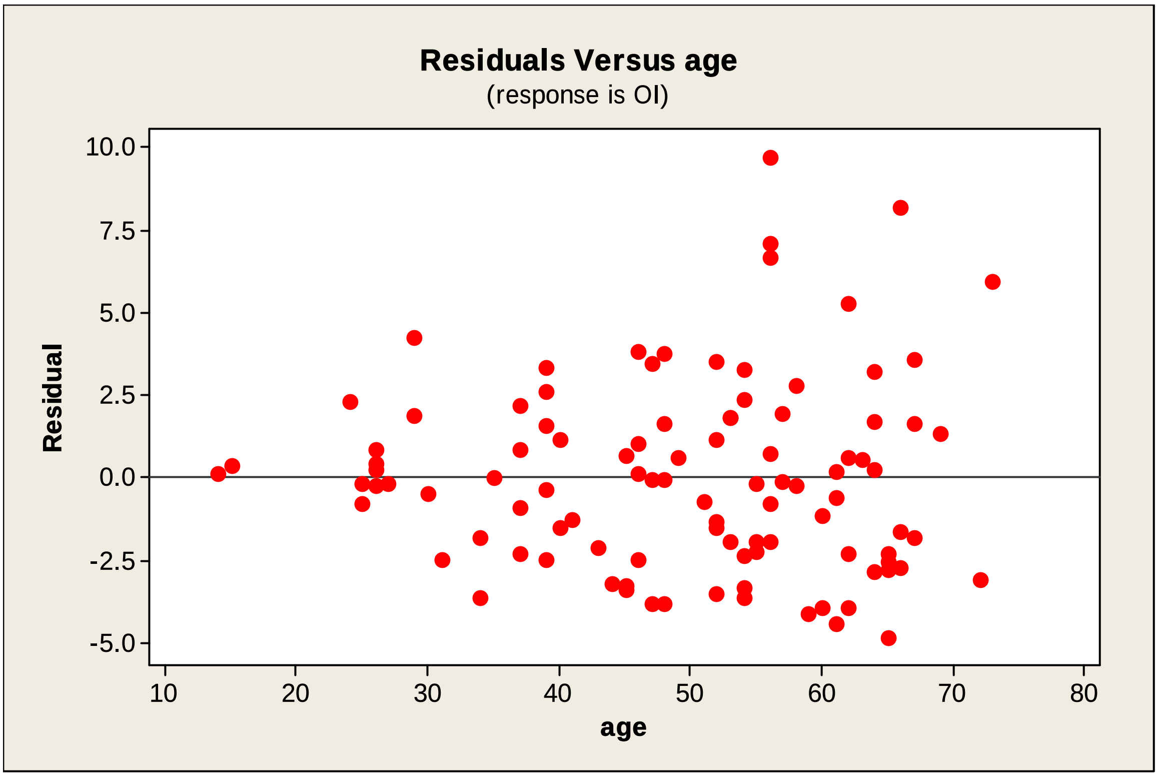
```
Predicted Values for New Observations

New
Obs     Fit  SE Fit       95% CI              95% PI
  1  0.7410  0.0204  (0.7005, 0.7815)  (0.3325, 1.1495)
```

The required interval is the one labelled 95% P I.

The problem with using OI, rather than log OI, in the analysis is that the assumption that individuals vary about the fitted line with constant SD is less tenable with the former variable than with the latter. To see this residuals need to be plotted against age for both regressions. To do this, the above procedure for fitting a regression line is followed but the Graphs… button is now clicked. In the dialogue box that appears, enter age in the Residuals versus the variables: box. Clicking on OK (twice) will recalculate the regression and plot the graph of residuals against age. The graph obtained from the regression on log OI is as follows



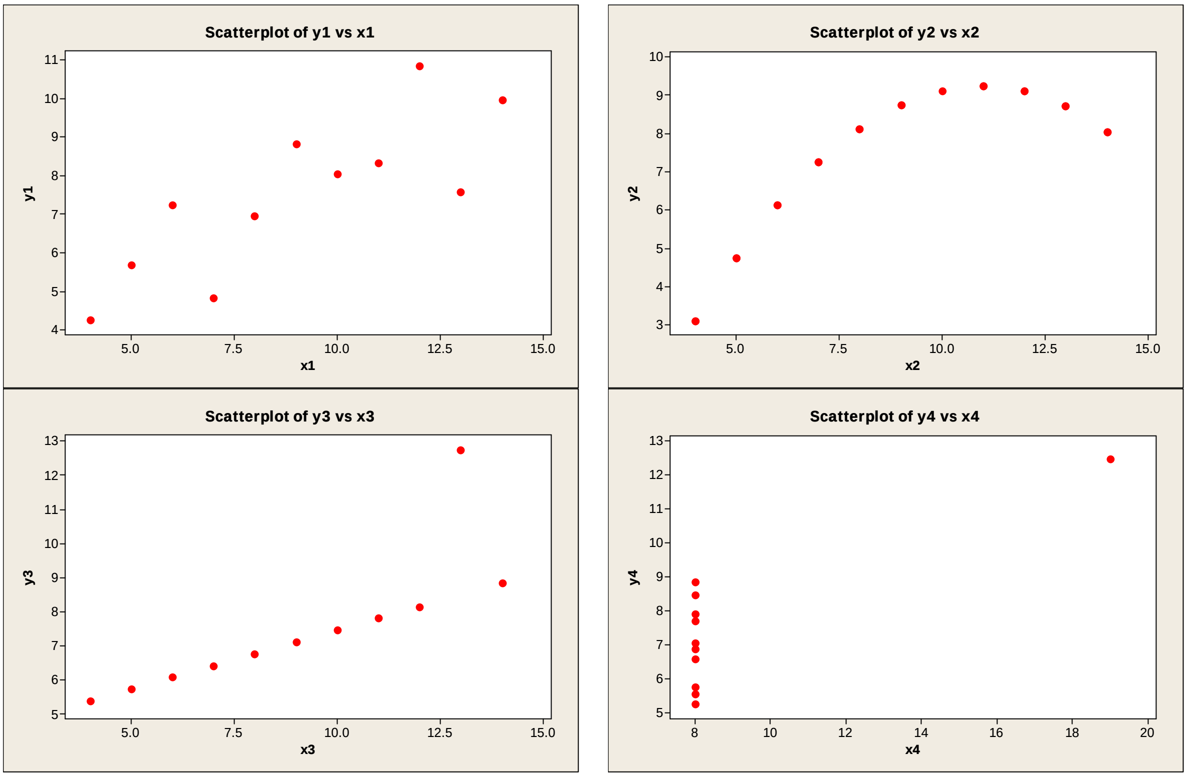
The corresponding plot when the Response: variable is OI is



Visual inspection suggests that the spread of the residuals is more or less the same at all ages in the upper graph but in the lower graph residuals at higher ages are more spread out than those at lower ages. This increase in spread with age violates the assumption that the SD about the line does not depend on age. It is for this reason that a regression of log OI on age is preferable to one of OI on age.

3.

The plots show very different patterns but all the correlations are the same (0.816). This shows the dangers of not plotting the data.


Scatterplot of y1 vs x1
y1
x1
Scatterplot of y2 vs x2
y2
x2
Scatterplot of y3 vs x3
y3
x3
Scatterplot of y4 vs x4
y4
x4