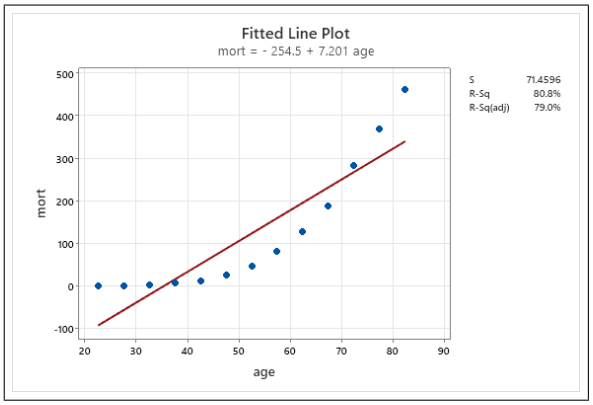
Additional Exercise 2.3

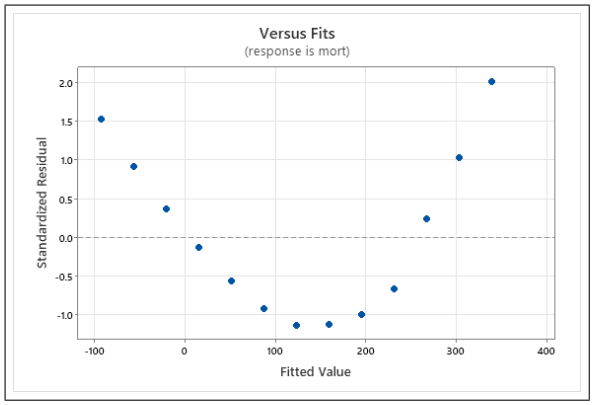
1. Comparison of simple linear regression models

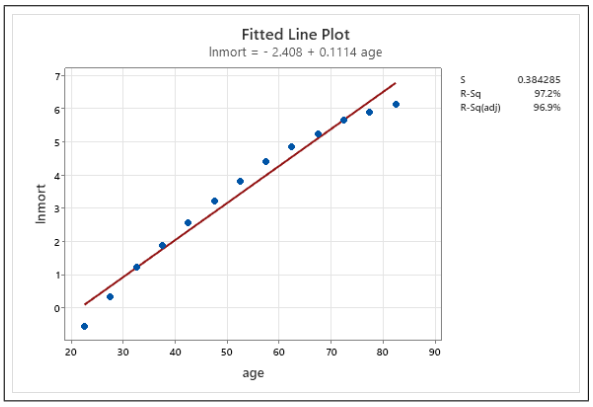
As described in the problem, we denote by y the colon cancer mortality for the different age groups (x). The three models described can be formulated,

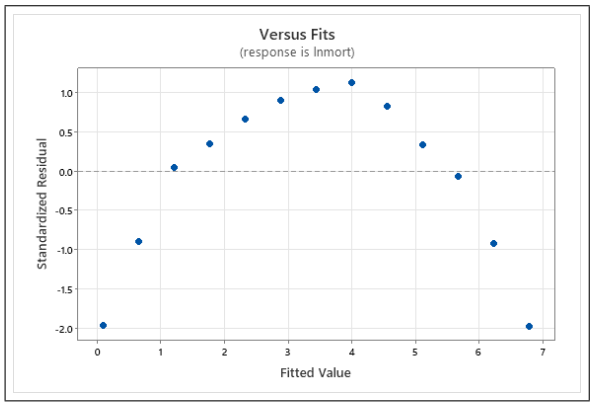
$$\begin{aligned} \text{(I):} \quad y_i &= \beta_0 + \beta_1\, x_i + \varepsilon_i, \\ \text{(II):} \quad \ln(y_i) &= \beta_0 + \beta_1\, x_i + \varepsilon_i, \\ \text{(III):} \quad \ln(y_i) &= \beta_0 + \beta_1\, \ln(x_i) + \varepsilon_i, \end{aligned}$$

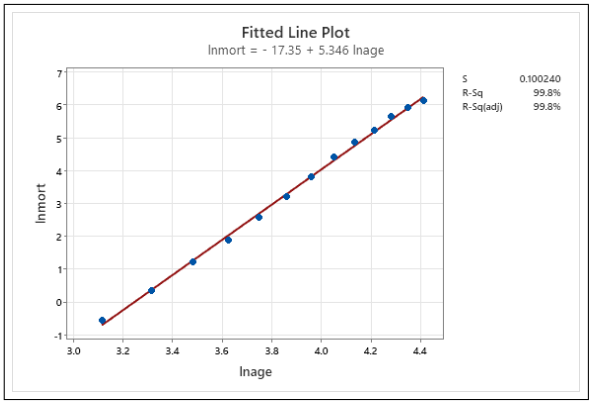
where in all models the errors $\varepsilon_1, \ldots, \varepsilon_{18}$ are assumed independent and identically distributed (i.i.d.) and normally distributed $\mathrm{N}(0, \sigma^2)$. The left part of the three sets of plots below, from the "Fitted Line Plot" menu in Minitab 21, gives the estimated regressions equations and shows the fitted curves overlaid the observed values. The right part shows the usual residual plot — the standardised residuals plotted against the fitted values.

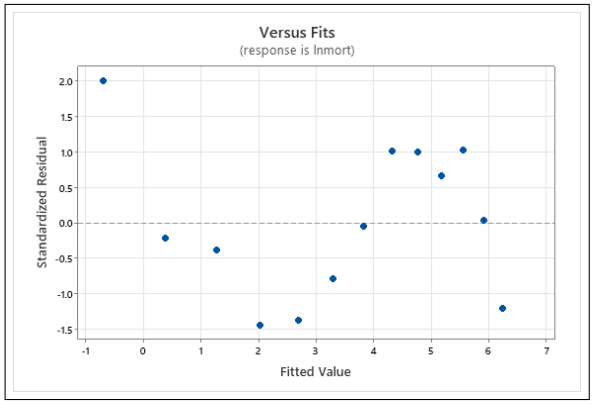












The three analyses and fitted line plots show clearly that the regression of log-mortality on log-age is the best, and the only one that would be somewhat acceptable. The two other models show an appreciable lack of fit. Even the log-log model has a clear pattern of the points around the line, which would indicate that a further refinement of the model is necessary. However, we will for this exercise accept the present fit as satisfactory. At $R^2 = 99.8\%$ the model already explains almost all of the variation in the data.

The fitted line plots already show the estimated regression line and residual standard deviation (s). For the log-log model we can backtransform the regression equation to

$$\hat{y} = \exp(-17.35) \cdot x^{5.346}.$$

The cancer mortality increases as a power function of x, and an estimated power of about 5.35. (Note that more decimals are needed to give precise predictions.) We include a table of predictions and prediction intervals (PI) for the two best models (Model I is not of real interest) for three selected age values.

	Observed	Model III		Model II	
age	y	$\hat{y}$	95% PI	$\hat{h}$	95% PI
27.5	1.42	1.45	$1.13 - 1.85$	1.93	$0.76 - 4.89$
52.5	45.7	45.9	$36.5 - 57.7$	31.2	$13.0 - 75.0$
77.5	369	368	$289 - 655$	504	$339 - 1280$

The difference between observed and fitted values is very small for the log-log model, and the prediction intervals are much narrower than those for the other model.

2. Cancel biology model

The proposed equation can be fitted by a multiple regression model, where the terms can be identified by comparing the equations:

$$\begin{aligned} \text{proposed:} \quad \ln(y_i) &= const + (n-1)\,\ln(x_i) - \lambda\, x_i + \varepsilon_i, \\ \text{multiple regression:} \quad \ln(y_i) &= \beta_0 + \beta_1\,\ln(x_i) + \beta_2\, x_i + \varepsilon_i. \end{aligned}$$

It is seen that the multiple regression model has the same age terms, and the parameters can be matched to the proposed equation by: $\beta_1 = (n-1)$ and $\beta_2 = -\lambda$. The Minitab listing from the multiple regression model is shown on the next page.

HS02_3.CSV

Regression Analysis: lnmort versus lnage, age

Regression Equation

lnmort = -16.19 + 4.927 lnage + 0.00897 age

Coefficients

Term	Coef	SE Coef	T-Value	P-Value
Constant	-16.19	1.11	-14.53	0.000
lnage	4.927	0.397	12.40	0.000
age	0.00897	0.00839	1.07	0.310

Model Summary

S	R-sq	R-sq(adj)
0.0995910	99.83%	99.79%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	2	57.9575	28.9788	2921.73	0.000
Error	10	0.0992	0.0099		
Total	12	58.0567			

Fits and Diagnostics for Unusual Observations

Obs	lnmort	Fit	Resid	Std Resid	
13	6.1356	6.2919	-0.1563	-2.08	R

R Large residual

The age-term has the wrong sign (λ should be > 0), but is clearly non-significant, so that there is no evidence against the regression coefficient being zero or slightly negative. The estimated value of λ is -0.009 (SE = 0.008). The estimated regression coefficient of log(age) is 4.9 (SE = 0.4) and strongly significant, therefore the estimated value of n is: $\hat{n} = 4.9 + 1 = 5.9$ or 6 mutating genes. Without the age-term, the estimated n would be $5.3 + 1 = 6.3$ or 6 as well.

The model still shows some lack of fit in the residual plot (not shown), but without knowledge about the real data (the data analysed here are aggregated data) it is difficult to improve the model.