**Solution to Exercise 12.2 (**Version 1, 14/8/15)

**from Statistical Methods in Biology: Design & Analysis of Experiments and Regression (2014) S.J. Welham, S.A. Gezan, S.J. Clark & A. Mead. Chapman & Hall/CRC Press, Boca Raton, Florida. ISBN: 978-1-4398-0878-8**



**Exercise 12.2** (Data: courtesy R. Harrington & C. Shortall, Rothamsted Research)

The Rothamsted Insect Survey collects insects using 12.2 m suction traps at locations across the UK. As part of an investigation into long-term changes in abundance of flying insects, indices of the total biomass collected per year (measured as wet weight) were created for the 30 years from 1973 to 2002 for four locations (Shortall et al., 2009). The wet weights (variate *WetWeight*, g) collected from the Hereford trap in each year (variate *Year*) are held in file HEREFORD.DAT. Use a SLR to investigate whether there is evidence of any linear trend over time in the log-transformed wet weights, calculated as $\log_{10}(WetWeight+0.5)$, and summarise the strength of the relationship. Use this model to predict the expected wet weight in 2010, and comment on the reliability of this prediction. Plot the fitted model and consider whether there are any aspects of the fit that you would wish to examine further. (We re-visit these data in Exercises 13.4 and 15.1.)

**Data 12.2 (HEREFORD.DAT)** Biomass of flying insects measured as wet weights (g) collected yearly in Hereford trap from 1973 to 2002.

| Year | WetWeight | Year | WetWeight | Year | WetWeight |
|---|---|---|---|---|---|
| 1973 | 2.480 | 1983 | 2.690 | 1993 | 1.000 |
| 1974 | 2.310 | 1984 | 0.960 | 1994 | 2.050 |
| 1975 | 1.620 | 1985 | 0.700 | 1995 | 1.380 |
| 1976 | 2.220 | 1986 | 2.480 | 1996 | 0.660 |
| 1977 | 1.660 | 1987 | 2.290 | 1997 | 0.650 |
| 1978 | 2.340 | 1988 | 2.460 | 1998 | 0.560 |
| 1979 | 1.870 | 1989 | 2.420 | 1999 | 0.400 |
| 1980 | 2.020 | 1990 | 0.750 | 2000 | 0.910 |
| 1981 | 2.890 | 1991 | 1.070 | 2001 | 1.000 |
| 1982 | 4.170 | 1992 | 1.210 | 2002 | 1.170 |

**Solution 12.2**

The data file contains two variates: *Year* and *WetWeight*. As instructed, we first take a log-transformation of the wet weights (g) as

$$logWt = \log_{10}(WetWeight + 0.5)$$

The offset of 0.5 matches that used in Shortall et al (2009). It is not strictly necessary in the context of this data set, as all of the weights are greater than 0.

First we examine the relationship graphically to check that an SLR model is plausible. The transformed wet weights are plotted against year number in Figure S12.2.1. There does appear to be an approximately linear downwards trend in the data, with variation about the line reasonably consistent across the sampling period. Fitting an SLR seems sensible here, so we will proceed.

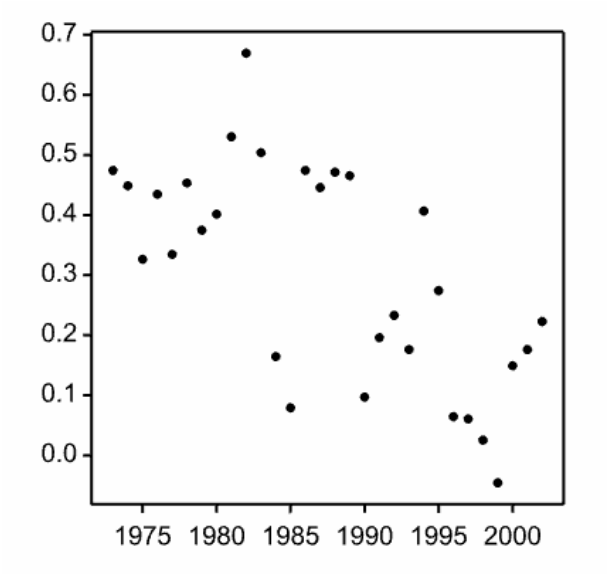


**Figure S12.2.1**. Log-transformed wet weights (g) plotted against sample year.

The SLR with year number as an explanatory variate is written in symbolic form as

Response: *logWt*
Explanatory component: *[1] + Year*

and in mathematical form as

$$logWt_i = \alpha + \beta Year_i + e_i$$

for $i = 1 \ldots 30$, using obvious variable names with the notation from Chapter 12. The summary ANOVA table from this model is Table S12.2.1. There is strong evidence of a linear relationship between logged wet weights and the explanatory variate *Year* ($F_{1,28} = 20.772$, $P < 0.001$), and the model accounts for 40.5% of the variation in the data (adjusted $R^2 = 0.405$). Before interpreting the model any further, we will check the residuals; a composite set of residual plots is in Figure S12.2.2. There is no suggestion of lack of fit to the linear trend, and the residuals seem consistent with a normal distribution with homogeneous variance. The fitted model has estimated slope $\hat{\beta} = -0.0134$ (SE 0.00295), indicating that the total biomass collected each year decreased during the survey period. Figure S12.2.3 shows the fitted model with the data and 95% confidence interval. As expected, the fitted model follows the downwards linear trend, but with substantial variation about the fitted line.

**Table S12.2.1** Summary ANOVA table for SLR with response $\log_{10}$(wet weight + 0.5) and year number as the explanatory variate.

| Source of variation | df | Sum of squares | Mean square | Variance ratio | *P* |
|---|---|---|---|---|---|
| Model | 1 | 0.4062 | 0.4062 | 20.772 | < 0.001 |
| Residual | 28 | 0.5475 | 0.0196 | | |
| Total | 29 | 0.9537 | 0.0329 | | |

**Table S12.2.2** Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (*P*) for a SLR model for logged wet weights with explanatory variate *Year*.

| Term | Parameter | Estimate | SE | t | *P* |
|---|---|---|---|---|---|
| *[1]* | α | 27.02 | 5.8624 | 4.609 | < 0.001 |
| *Year* | β | -0.01344 | 0.002950 | -4.558 | < 0.001 |

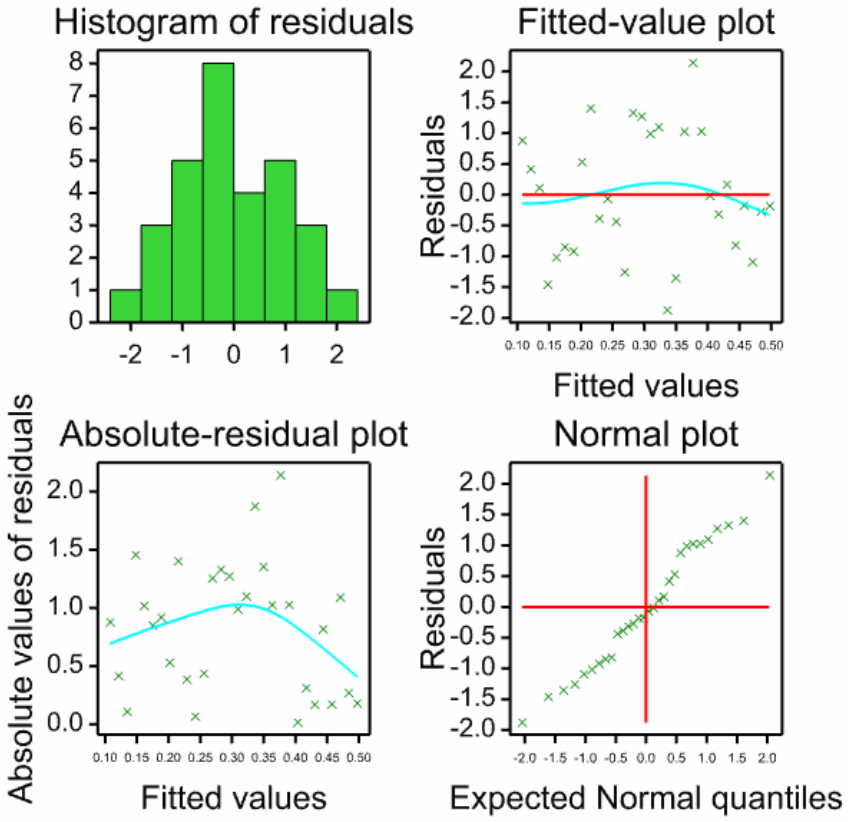


**Figure S12.2.2.** Composite set of residual plots from SLR for log-transformed wet weights with year number as the explanatory variate.

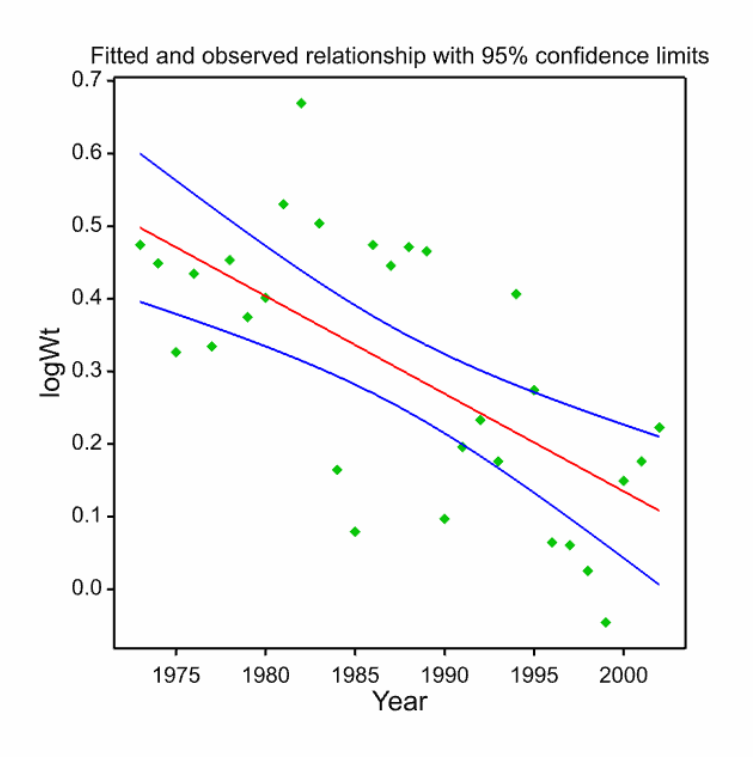


**Figure S12.2.3**. Fitted SLR with 95% confidence interval.

The predictive model (on the log-scale) is

$$\hat{\mu}(Year) = 27.02 - 0.0134\,Year$$

We can get a prediction from this fitted SLR for 2010 as 0.00 (= 27.02 − 0.0134×2010) with SE 0.0711 and 95% CI (-0.145, 0.146). These predictions are on the log-scale, and we can back-transform them into predicted wet weights ($\hat{W}$) as

$$\hat{W}(Year) = 10^{\hat{\mu}(Year)} - 0.5\,.$$

The back-transformed predicted wet weight is 0.501 units, with 95% confidence interval (0.216, 0.900). However, we should remember that this prediction is an extrapolation that assumes that the average linear trend (on the log-scale) observed over the 30-year period would continue for another 8 years – there is no justification for this assumption, and this prediction is unlikely to be reliable.

In terms of checking the model assumptions, the plot of the fitted model shows no suggestion of model misspecification or variance heterogeneity. However, as the measurements are made at the same place over time, it would be sensible to check for any sign of temporal correlation, for example by plotting the residuals against year number, or by plotting each residual against that from the previous year (see Section 5.5.2 for more details). Figure S12.2.4 includes both types of plot. There is a suggestion of some positive correlation between successive residuals, calculated as correlation = 0.34. For a simple model like SLR with a clear result, we might choose to ignore this correlation. For more complex models, or with less clear-cut results, we should try to account for the correlation to avoid any misleading results; e.g, we might impose a correlation structure over time using linear mixed models.

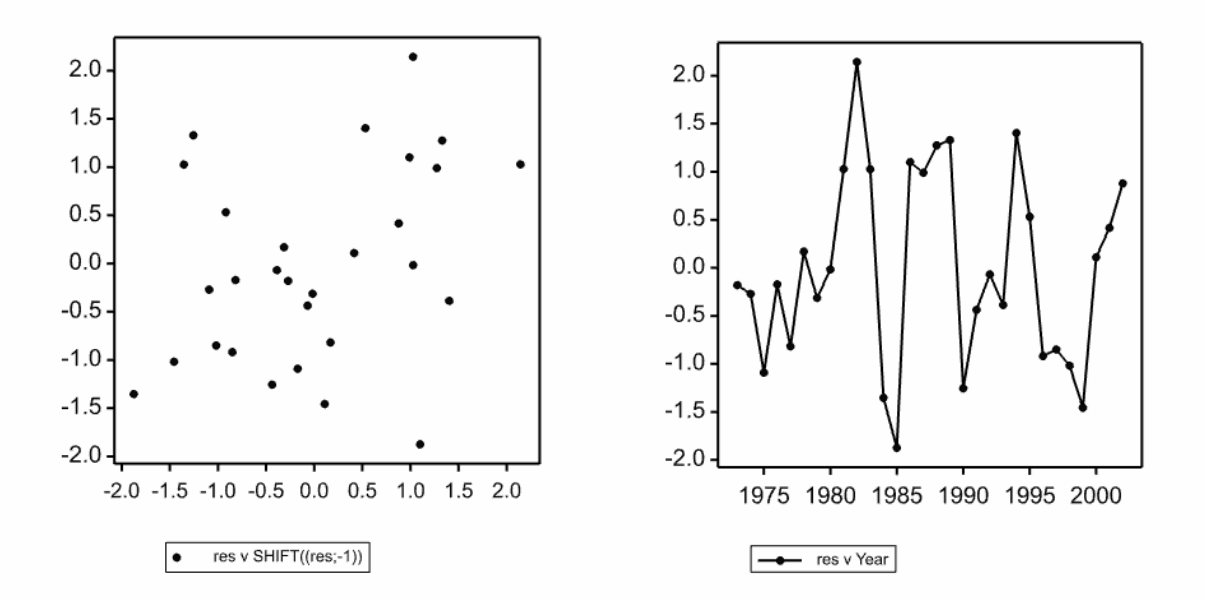


**Figure S12.2.4**. Left: residuals plotted against that from previous year. Right: residuals from SLR plotted against year number.