Solution to Exercise 13.2 (Version 1, 6/12/15)

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 13.2*

Now consider the original data from the experiment described in Exercise 12.1. The numbers of leaves on each plant (variate *NLeaves*) are in file CABBAGE.DAT with unit numbers (*ID*) and sample dates (variate *Days*). Fit a SLR and use diagnostic plots to check the fit of the model. Would a transformation be appropriate here? If so, implement it and re-fit the SLR on your chosen scale. Plot the fitted model and check for any evidence of lack of fit. Give a 95% CI for the growth rate over the period (as leaves per day) and interpret this estimate. Can you reconcile this result with the one you gave in Exercise 12.1? (We re-visit these data in Exercise 18.2.)

Data 13.2 (CABBAGE.DAT)

ID	Days	NLeaves
1	0	10
2	0	6
3	0	9
4	0	7
5	8	10
6	8	7
7	8	11
8	8	11
9	14	10
10	14	7
11	14	9
12	14	12
13	21	5
14	21	13
15	21	12
16	21	14
17	28	14
18	28	23
19	28	13
20	28	14
21	35	20
22	35	21
23	35	12
24	35	16
25	37	14
26	37	13
27	37	23
28	37	20
29	42	14
30	42	18
31	42	16
32	42	29
33	44	20
34	44	20
35	44	18
36	44	16
37	46	13
38	46	17
39	46	20
40	46	21

Solution 13.2

The numbers of leaves are plotted against sample dates in Figure S13.2.1. The number of leaves is larger for larger numbers of days after planting – it is not clear that the increase is linear, but this is plausible so we fit the SLR as instructed. The model for the SLR can be written symbolically as

Response:	*NLeaves*
Explanatory component:	*[1]* + *Days*

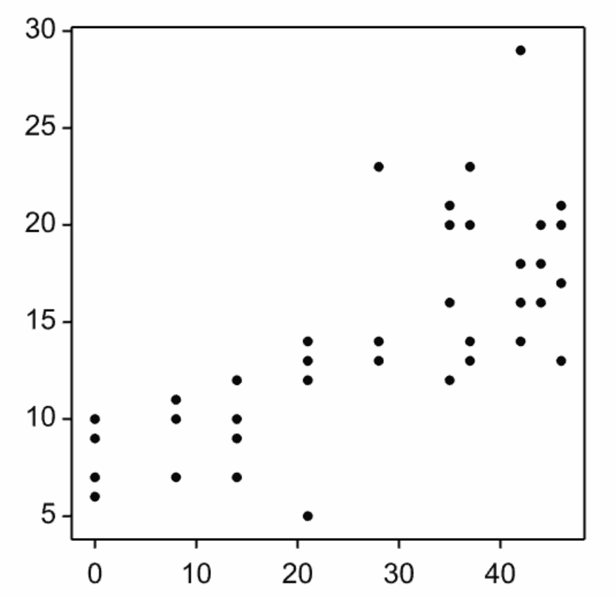

Figure S13.2.1. Number of leaves plotted against days after planting.

Table S13.2.1 Summary ANOVA table for SLR with response number of leaves and sample date (days after planting) as the explanatory variate.

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*
Model	1	630.81	630.81	47.647	< 0.001
Residual	38	503.09	13.24		
Total	39	1133.90			

Table S13.2.2 Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (*P*) for a SLR model for number of leaves (*NLeaves*) with explanatory variate sample date (days after planting, *Days*).

Term	Parameter	Estimate	SE	t	*P*
[1]	α	7.30	1.185	6.160	< 0.001
Day	β	0.26	0.038	6.903	< 0.001

The summary ANOVA table from the SLR is Table S13.2.1. There is strong evidence ($F_{1,38}$ = 46.65, $P < 0.001$) of a linear relationship between the number of leaves present and number of days after planting, which accounts for 54.5% of the variation (adjusted $R^2 = 0.545$). The estimated parameters of the fitted SLR are shown in Table S13.2.2. Figure S13.2.2 shows (a) the fitted model and (b) the residual plotted against the explanatory variate. There is no suggestion of model misspecification, as there is no indication of curvature about the fitted line. The spread of observations about the fitted line does appear larger for larger fitted values (or larger numbers of days). We can investigate this further using a composite set of residual plots, as shown in Figure S13.2.3. The histogram of residuals appears slightly skewed, and the fitted value and absolute residual plots again suggest some variance heterogeneity.

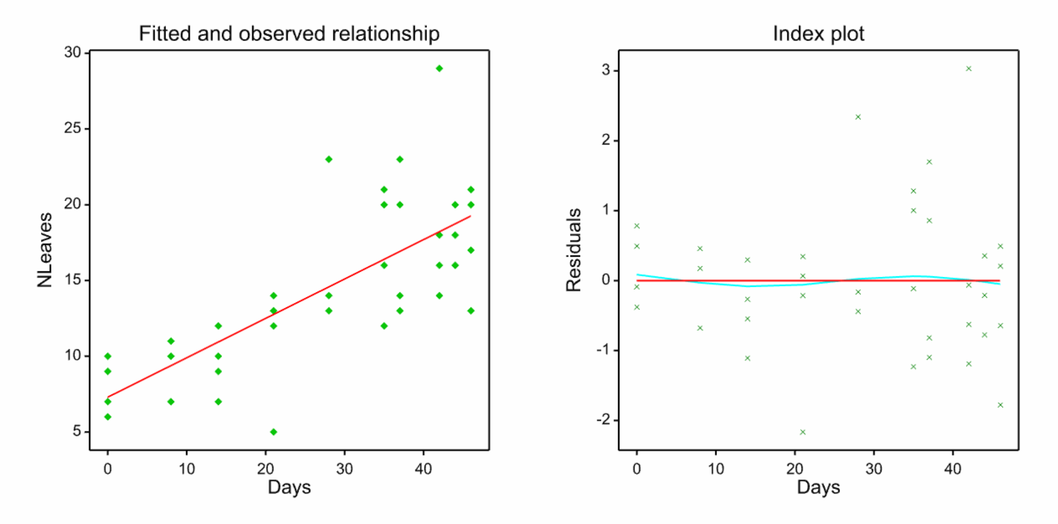

Figure S13.2.2. (a) Observed number of leaves with fitted SLR; (b) plot of residuals from SLR against explanatory variate.

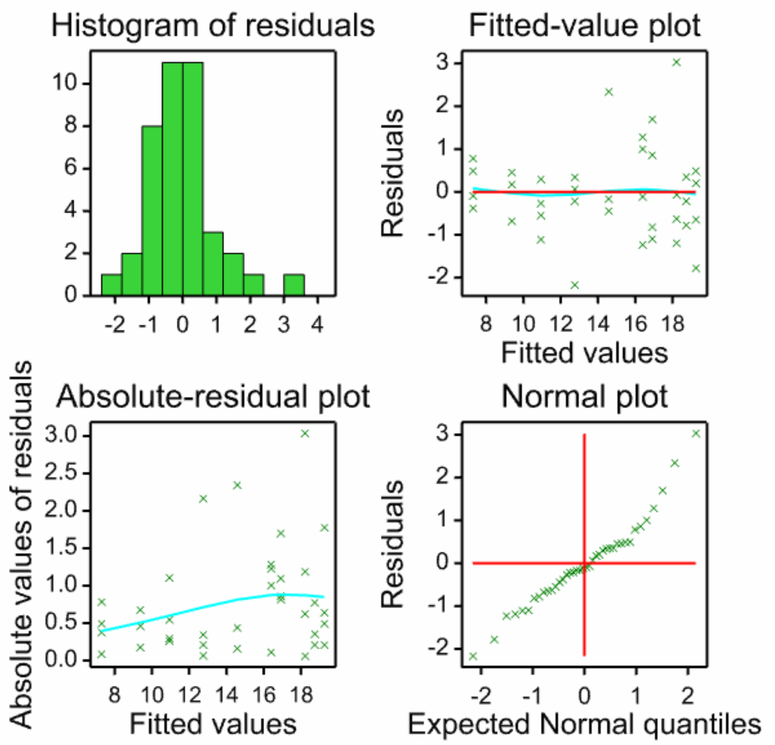

Figure S13.2.3. Composite set of residual plots (based on standardized residuals) from SLR for number of leaves.

The residual plots are consistent with the presence of variance heterogeneity, suggesting larger residual variance for larger fitted values. We will try a transformation to address this issue. Because the number of leaves is a count, and a logarithm transformation often works well for count variables, we will first investigate a logarithm transformation. We calculate the natural logarithm transformation as

$$logNL = \log(NLeaves).$$

A plot of the transformed response plotted against number of days after planting is shown in Figure S13.2.4. The spread of observations for each number of days about their mean appears more equal than in Figure S13.2.1, and the relationship still appears to be plausibly linear, so will we proceed to fit an SLR model to this new response variable, with symbolic form

Response:	*logNL*
Explanatory component:	*[1]* + *Days*

The summary ANOVA table from this SLR is Table S13.2.4. There is again strong evidence ($F_{1,38}$ = 56.12, $P < 0.001$) of a linear relationship, here between the log number of leaves and number of days after planting. This model accounts for 58.6% of the variation on the log scale (adjusted $R^2 = 0.586$). The estimated parameters of the fitted SLR are shown in Table S13.2.5. Figure S13.2.5 shows (a) the fitted model and (b) standardized residuals plotted against the explanatory variate. There is again no suggestion of model misspecification, and the spread of observations about the fitted line appears similar across the full range. The composite set of residual plots from this model, shown in Figure S13.2.6, appears satisfactory except for one possible outlying observation (with number of days = 20).

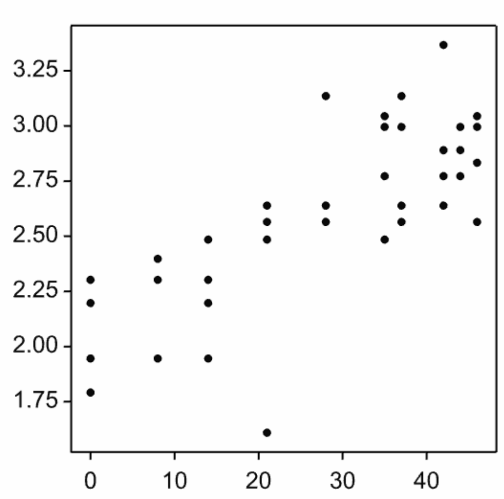

Figure S13.2.4. Log(number of leaves) plotted against sample date (days after planting).

Table S13.2.4 Summary ANOVA table for SLR with response log(number of leaves) and sample date (days after planting) as the explanatory variate.

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*
Model	1	3.753	3.753	56.118	< 0.001
Residual	38	2.541	0.067		
Total	39	6.294			

Table S13.2.5 Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (*P*) for a SLR model for log(number of leaves) (*logNL*) with explanatory variate sample date (days after planting, *Days*).

Term	Parameter	Estimate	SE	t	*P*
[1]	α	2.046	0.0842	24.289	< 0.001
Day	β	0.020	0.0027	7.491	< 0.001

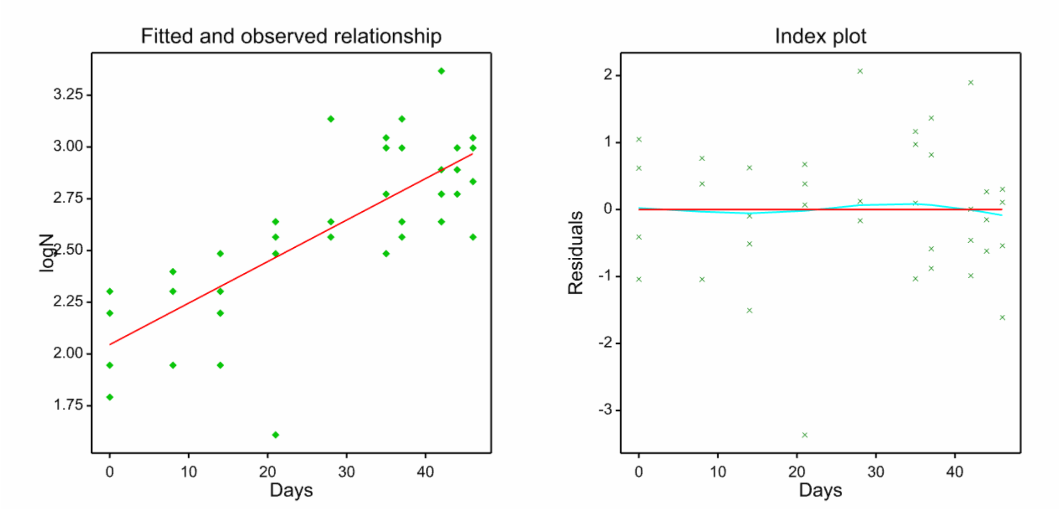

Figure S13.2.5. (a) Observed number of leaves with fitted SLR for log-transformed leaves; (b) plot of residuals from SLR for log-transformed leaves against explanatory variate.

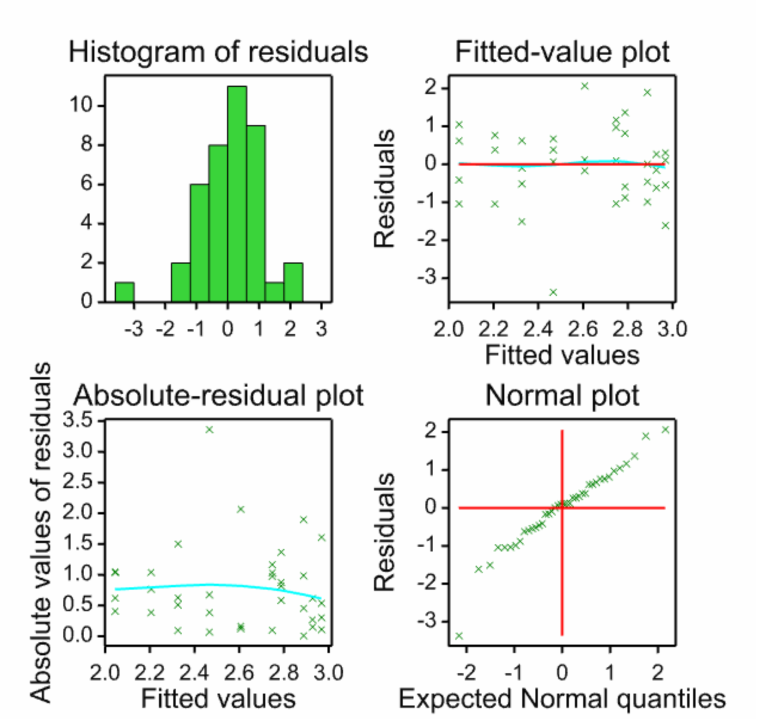

Figure S13.2.6. Composite set of residual plots (based on standardized residuals) for SLR with response log(number of leaves).

We might investigate whether there is any question of an error in the smallest observation at 20 days, but the discrepancy does not look abnormal on the untransformed scale so we would not omit the observation unless we have additional evidence that it is incorrect. On balance, the SLR on the transformed scale appears to satisfy our assumptions of a Normal distribution with equal variance for the model deviations more closely than the SLR with the untransformed number of leaves. We will proceed to make a formal check for lack of fit, by adding a factor version of the explanatory variate to the model. We call the factor version of the explanatory variate fDays, and the model testing for lack of fit can be written in symbolic form as

Response:	*logNL*
Explanatory component:	*[1]* + *Days* + fDays

The sequential ANOVA table for this model is Table S13.2.6. The change in SS associated with the factor fDays measures lack of fit in terms of variation of group means (with one group for each number of days) about the fitted regression line. For this model, there is no evidence of any lack of fit to the SLR model on the log-transformed scale ($F_{8,30}$ = 0.479, *P* = 0.861). We therefore accept the SLR fitted to log-transformed number of leaves, with the estimated parameters shown in Table S13.2.5.

Table S13.2.6. Sequential ANOVA table for SLR with lack-of-fit for log(number of leaves) with days after planting as the explanatory variate.

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*
+ *Days*	1	3.753	3.753	49.968	< 0.001
+ fDays	8	0.288	0.036	0.479	0.861
Residual	30	2.253	0.075		
Total	39	6.294	0.161		

In mathematical terms, we can write the fitted model as

$$\log(N_{ij}) \;=\; \hat{\alpha} + \hat{\beta}\,Days_{ij}\,,$$

where N_{ij} and $Days_{ij}$ are the number of leaves and number of days after planting, respectively, for the j^{th} observation in the i^{th} group; the subscript *i* ranges from 1 to 10 and *j* from 1 to 4. We can back-transform this to write the model in terms of the predicted number of leaves as

$$\hat{N}_{ij} \;=\; \exp(\,\hat{\alpha} + \hat{\beta}\,Days_{ij}\,)\,.$$

We can rewrite this model in generic form for prediction for any number of days (*Days*, preferably within the range 0-46) as

$$\hat{N}(Days) \;=\; \exp(\,\hat{\alpha} + \hat{\beta}\,Days\,)\,.$$

This is no longer a linear model, so we can no longer use the slope as an estimate of the growth rate, as we could in an SLR for the original numbers of leaves. The growth rate in this model is estimated by taking the derivative of the model with respect to the *Days* variable, as

$$\frac{\partial \hat{N}(Days)}{\partial Days} = \hat{\beta}\exp(\,\hat{\alpha} + \hat{\beta}\,Days\,),$$

ie. the predicted value (on the back-transformed scale) multiplied by the slope of the SLR (on the log scale). As the slope estimate ($\hat{\beta} = 0.020$) is positive and the predicted values are all positive, then the estimated growth rate is always positive. Within the observed range, the growth rate takes its minimum value of 0.155 leaves per day at day 0 and its maximum value of 0.390 leaves per day at day 46. Calculating a 95% CI for the growth rate is not straightforward, as we have a non-linear function of the estimated parameters, but this can be achieved using the so-called delta method, available in most statistical packages. We thus find 95% CI's for growth rate as (0.132, 0.178) at day 0 and (0.242, 0.538) at day 46.

In Exercise 12.1, an SLR was fitted directly to the mean number of leaves at each sample date, with estimated intercept equal to 7.299 and estimated slope equal to 0.261. This slope estimate is in the middle of the range of growth rates found for our new model, and the predicted number of leaves for day 0 from the new model is 7.734 (compared to observed mean of 8.0 leaves). In contrast to our findings here, there was no evidence of variance heterogeneity in the residual plots from the SLR in Exercise 12.1. There are several possible reasons for this:

- the mean numbers gave a smaller data set, with less information available to investigate heterogeneity
- the mean numbers had no replication, so it was not possible to separate deviation from lack-of-fit to the proposed model, making it harder to detect heterogeneity
- by chance, we are seeing heterogeneity in the raw numbers that is not actually present

The last option seems the least likely, given that the data are counts (where we often observe heterogeneity) and the observed variation is larger when the mean number of leaves is larger (again as we would expect for counts), but we cannot be certain. In practice, predictions from the two models are very similar.