Solution to Exercise 12.6 (Version 1, 15/8/15)

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 12.6 (Data: courtesy V. Buchanan-Wollaston (PRESTA), University of Warwick)

A microarray study investigated genes associated with the senescence of leaves. Forty-four plants were grown in a controlled environment and the seventh leaf was excised from four of these plants at two-day intervals from 19–39 days after sowing (at the same point in the day/night cycle each time). The plants were allocated to sample dates at random, with a CRD design. Four sub-samples (technical replicates) were taken from each leaf and allocated to separate microarrays. File SENESCENCE.DAT holds unit numbers (*ID*), design information (variate *Day*, factor BiolRep) and the expression value for three genes (variates *CATMA3A13560*, *CATMA2A31585* and *CATMA1A09000*) from each plant following normalization and combination of the values for the four technical replicates. Use SLR to predict the expression of gene *CATMA3A13560* over time. Is there any evidence of lack of fit to this relationship?

Data 12.6 (SENESCENCE.DAT). Day of measurement and biological replicate (BR) for genes Catma3A13560 (C3A), Catma2A31585 (C2A) and Catma1A09000 (C1A).

ID	Day	BR	C3A	C2A	C1A	ID	Day	BR	C3A	C2A	C1A
1	19	1	6.4053	12.8395	9.1348	23	19	3	6.5969	13.2640	8.6787
2	21	1	7.9146	12.2470	8.9780	24	21	3	7.5449	12.4184	8.9637
3	23	1	7.7789	11.7455	9.5849	25	23	3	8.2989	12.1062	9.1097
4	25	1	8.8533	11.8965	9.3157	26	25	3	8.7385	11.7594	9.4909
5	27	1	8.3601	11.9451	9.6473	27	27	3	8.6489	11.4728	10.6307
6	29	1	8.9138	11.8739	10.7223	28	29	3	9.0198	11.6767	10.3833
7	31	1	10.6223	11.6587	11.2268	29	31	3	10.2474	11.9051	10.6632
8	33	1	10.3948	11.6235	11.1481	30	33	3	10.8396	11.7603	11.5779
9	35	1	9.4488	11.7293	11.2526	31	35	3	10.7966	11.7872	11.4960
10	37	1	10.5779	11.3027	11.7078	32	37	3	11.8950	9.8893	12.3864
11	39	1	10.4757	11.4350	10.9402	33	39	3	10.4576	11.1802	11.4699
12	19	2	8.0890	12.6862	9.2724	34	19	4	7.9906	12.6954	8.8843
13	21	2	7.7509	12.9831	8.7144	35	21	4	9.3149	12.5664	8.7321
14	23	2	6.6537	12.2413	8.7275	36	23	4	7.2174	11.6354	8.9985
15	25	2	8.8790	11.8709	9.2369	37	25	4	8.2152	11.7469	9.8572
16	27	2	7.9490	11.6430	10.0873	38	27	4	8.4626	11.4095	10.0430
17	29	2	8.8540	11.9500	10.4858	39	29	4	9.7709	12.2218	10.8605
18	31	2	9.9089	11.8449	10.3783	40	31	4	10.1883	11.7061	10.8431
19	33	2	11.0003	11.2225	12.1593	41	33	4	11.2479	11.6132	11.7492
20	35	2	9.9794	11.3624	10.8485	42	35	4	10.0495	11.6094	11.3632
21	37	2	10.1501	11.1616	11.3656	43	37	4	9.7399	11.7092	10.9349
22	39	2	11.2561	10.7493	11.4014	44	39	4	10.5279	10.9310	11.2358

Solution 12.6

Here we are interested in the expression values for gene Catma3A13560, held in variate *CATMA3A13560*. We start by plotting the expression values against the explanatory variate *Day* to see if there is any indication of a relationship, as in Figure S12.6.1. The relationship appears approximately linear, if a bit noisy. We proceed with fitting a SLR.

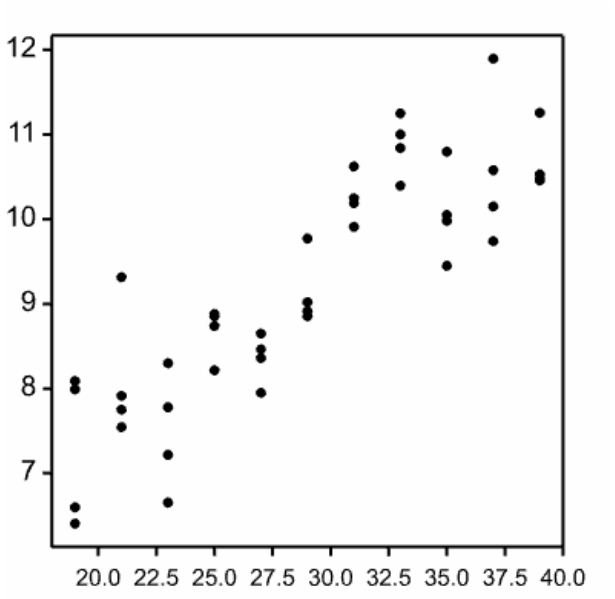

Figure S12.6.1. Expression values for gene Catma3A13560 plotted against day of excision.

The symbolic form for the SLR model is

Response: *CATMA3A13560*
Explanatory component: *[1] + Day*

The summary ANOVA table for this model is Table S12.6.1. There is very strong evidence of a linear relationship between day of leaf excision and the gene expression measurement ($F_{1,42} = 122.50$, $P < 0.001$). The SLR accounts for 73.9% of the variation in expression values (adjusted $R^2 = 0.739$).

Table S12.6.1 Summary ANOVA table for SLR with response expression for gene Catma3A13560 and day of leaf excision as the explanatory variate.

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*
Model	1	60.8619	60.8619	122.495	< 0.001
Residual	42	20.8679	0.4969		
Total	43	81.7297			

Table S12.6.2 Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (*P*) for a SLR model for expression of gene Catma3A13560 (variate *CATMA3A13560*) with explanatory variate *Day*.

Term	Parameter	Estimate	SE	t	*P*
[1]	α	3.835	0.4987	7.690	< 0.001
Day	β	0.186	0.0168	11.068	< 0.001

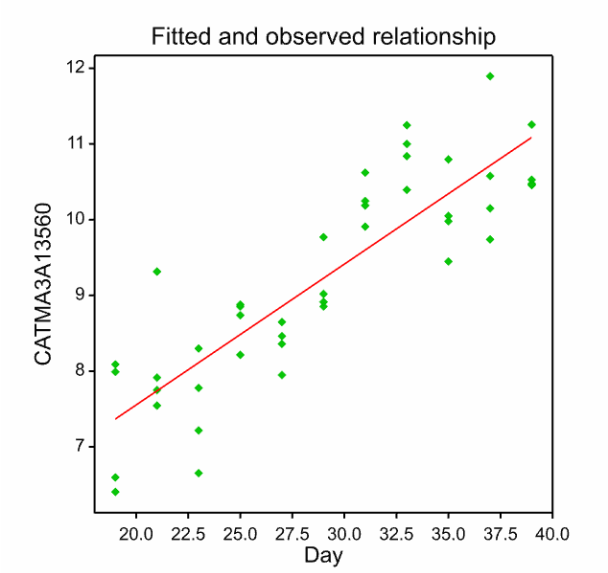

Figure S12.6.2. Fitted SLR with observed data for expression of gene Catma3A13560 with day of leaf excision as explanatory variate.

The estimated parameters are in Table S12.6.2. We estimate that gene expression measurements increase by 0.186 units for each additional day before leaf excision. A plot of the fitted model (Figure S12.6.2) shows that the fitted line runs mainly through the spread of the data, although the data for some sample dates lie either entirely above (31, 33 days) or below (27 days) the fitted line. This type of pattern would be expected if there was some curvature, or other lack of fit, to the SLR model. Figure S12.6.3 shows a composite set of residual plots from this SLR. The fitted values plot again shows consistent deviation about the fitted line at 27, 31 and 33 days. The residuals otherwise seem consistent with a Normal distribution with common variance, and so the assumptions underlying the model appear to be satisfied.

As we have replicate observations at each sample date, we can make a formal test for lack of fit to the SLR model. We do this by creating a factor with a separate level for each of the sampling days, we call this factor *fDay*. We then fit a model with symbolic form

Response: *CATMA3A13560*
Explanatory component: *[1] + Day + fDay*

and examine the sequential ANOVA table associated with this model (Table S12.6.3).

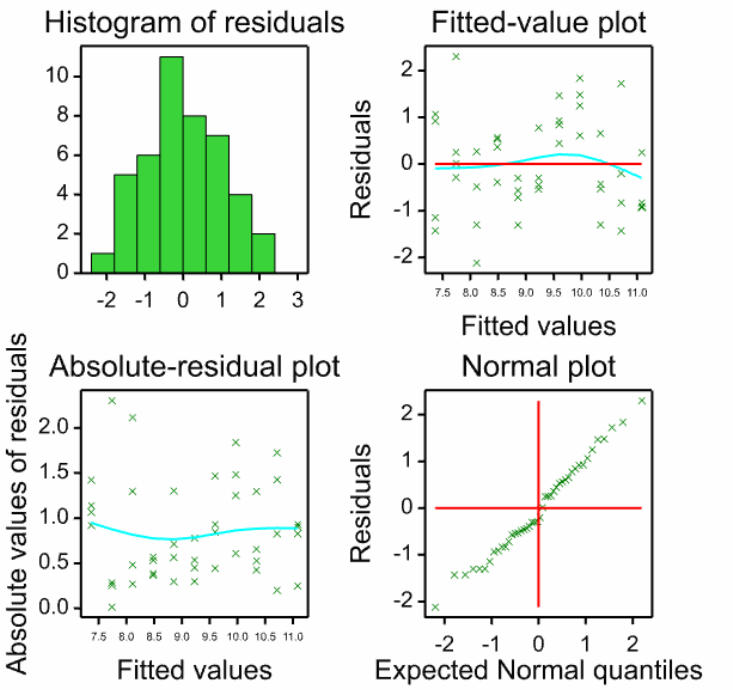

Figure S12.6.3. Residual plots based on standardized residuals from SLR for expression of gene Catma3A13560 with day of leaf excision as the explanatory variate.

Table S12.6.3 Sequential ANOVA table for SLR with lack-of-fit for expression for gene Catma3A13560 with day of leaf excision as the explanatory variate.

Source of variation	df	Sum of squares	Mean square	Variance ratio	*P*
+ *Day*	1	60.8623	60.8623	173.61	< 0.001
+ fDay	9	9.2990	1.0332	2.95	0.011
Residual	33	11.5687	0.3506		
Total	43	81.7297			

The sum of squares (SS) associated with factor fDay represents lack of fit (LOF) to the SLR model, and the residual SS can now be interpreted as pure error (background variation between units with the same treatment). In this model, there is evidence for the presence of LOF ($F_{9,33} = 2.95$, $P = 0.011$), indicating that the deviations away from the fitted line cannot reasonably be attributed to random background variation. We can examine the pattern of LOF by comparing the fitted values from the two models, as shown in Figure S12.6.4. There is no indication of smooth trend in the LOF, so no suggestion that we should be fitting a different form of model (such as a curved or non-linear model). If there were regular cycles in the expression values with a frequency that was quite short relative to the two-day sampling intervals, then we would not see many observations per cycle, and this could lead to this sort of "up and

down" scatter about the straight line. Some knowledge of the biological system is required to say whether this might be plausible, and further research, using smaller lags between samples, would be necessary to demonstrate it. However, many other explanations are possible, and the fitted SLR certainly provides a reasonable guide as to the overall level of expression over time, even if it does not tell the full story.

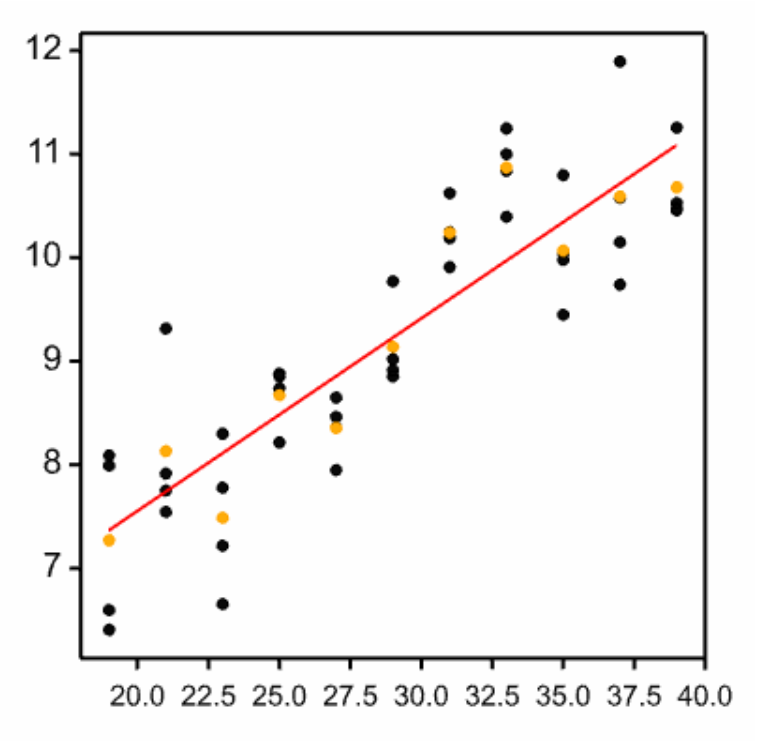

Figure S12.6.4. Observed expression of gene Catma3A13560 (black dots) with fitted SLR (red line) and fitted group means (orange dots).