Solution to Exercise 12.4 (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 12.4 (Data: courtesy I. Shield, Rothamsted Research)

An experiment was established to identify varieties of willow with high yields of dry matter. However, as accurate measurement of dry matter is time-consuming, use of a surrogate variable is desirable and several such variables were measured on a sample of 113 trees. File WILLOWSTEMS.DAT holds the values of dry matter (variate *DryMatter*) and several summary variables, including the length of the longest stem (variate *MaxLength*), which is the simplest to measure. Fit a SLR relating dry matter to the maximum stem length — could we reasonably use this as a surrogate variable? (We re-visit these data in Exercises 13.5 and 14.5.)

Data 12.4 (WILLOWSTEMS.DAT)

Properties of a stems from a sample of 113 willow trees: ID = sample number, DM = total dry matter (DryMatter), SL = sum of stem lengths (SumLength), SD = sum of stem diameters (SumDiam), ML = maximum stem length (MaxLength), LT5 = length of longest 5 stems (LengthTop5).

ID	DM	SL	SD	ML	LT5	ID	DM	SL	SD	ML	LT5
1	1.22	1947	78	137	115.8	58	15.04	6349	261	502	481.0
2	3.99	2728	169	241	213.4	59	15.11	5751	207	553	498.0
3	5.37	3527	126	295	271.6	60	15.20	5967	211	410	357.6
4	5.67	1609	59	283	203.8	61	15.53	7241	286	423	412.4
5	6.13	2502	129	280	231.8	62	15.56	6797	206	373	349.2
6	6.76	2877	123	327	299.0	63	15.67	8532	349	462	439.0
7	6.84	4058	191	449	406.4	64	15.91	5826	281	472	389.4
8	6.91	2415	93	363	306.6	65	15.92	6631	261	544	509.4
9	7.49	3459	140	368	347.0	66	15.93	9620	352	425	412.4
10	8.29	151	4	151	151.0	67	16.01	5958	214	448	396.4
11	8.50	2979	111	383	361.8	68	16.05	3829	166	524	466.4
12	9.10	4421	176	376	351.6	69	16.18	10966	372	405	396.0
13	9.29	4808	173	357	328.6	70	16.28	9609	346	430	419.2
14	9.62	2110	93	378	343.4	71	16.28	9181	344	520	471.4
15	9.81	7605	300	363	332.8	72	16.65	7826	343	475	448.2
16	9.90	7029	289	391	351.4	73	16.86	6475	255	540	517.8
17	10.43	6353	234	364	333.0	74	16.96	5631	234	396	368.0
18	10.74	8948	360	296	289.2	75	17.15	7506	283	471	453.2
19	10.80	8097	278	361	338.0	76	17.15	8106	327	400	377.2
20	11.05	2718	106	526	393.8	77	17.22	5554	216	471	456.2

ID	DM	SL	SD	ML	LT5	ID	DM	SL	SD	ML	LT5
21	11.21	5462	261	309	299.0	78	17.27	9448	340	448	418.0
22	11.45	7697	250	416	383.4	79	17.36	10067	332	378	371.4
23	11.51	4572	173	449	399.2	80	17.49	6800	367	441	407.4
24	12.16	6332	257	341	313.2	81	17.54	6742	281	526	482.2
25	12.25	6546	235	366	354.8	82	17.65	7960	286	343	328.2
26	12.34	12958	368	342	316.6	83	17.84	8853	369	393	353.0
27	12.56	11477	293	340	327.0	84	17.96	10440	355	421	403.0
28	12.67	8060	344	296	282.0	85	18.02	9223	331	464	408.0
29	12.68	5589	205	366	352.2	86	18.28	7159	274	417	393.8
30	12.78	10615	376	391	363.8	87	18.44	5855	231	519	462.4
31	12.85	3966	170	510	473.6	88	18.52	5536	217	507	483.4
32	12.89	6196	210	447	407.2	89	18.96	7768	275	472	433.2
33	12.97	5581	225	395	360.8	90	19.02	5517	213	571	512.0
34	13.01	7164	293	435	419.4	91	19.12	10513	333	466	429.6
35	13.01	5947	217	447	396.6	92	19.16	6589	246	449	421.8
36	13.13	7127	252	353	328.8	93	19.30	8844	328	383	371.6
37	13.22	4622	191	472	412.8	94	19.50	7892	325	483	408.8
38	13.24	6554	276	444	406.6	95	19.61	8502	430	425	415.2
39	13.30	6491	259	453	436.6	96	19.70	5005	154	589	490.0
40	13.33	5987	242	432	413.2	97	19.80	8300	310	388	374.0
41	13.40	4384	194	409	358.6	98	19.99	8380	352	427	397.0
42	13.54	6730	257	425	372.2	99	20.25	6638	251	410	396.0
43	13.61	7153	263	388	368.0	100	20.46	6501	260	509	481.6
44	13.73	6886	294	391	364.8	101	20.47	7118	280	562	539.2
45	13.75	5809	214	402	384.4	102	20.52	5585	226	521	508.6
46	13.76	5850	240	326	315.6	103	20.90	12664	416	356	347.4
47	13.82	4941	228	450	429.6	104	20.94	6211	313	544	527.6
48	13.82	7603	338	360	336.6	105	20.98	5840	231	511	468.0
49	14.00	7552	352	384	346.0	106	21.28	7367	288	499	488.6
50	14.28	7158	211	447	420.2	107	21.92	8976	435	420	409.6
51	14.40	6439	252	423	387.8	108	22.02	5291	229	477	431.6
52	14.41	6416	243	419	404.6	109	22.30	9357	373	457	443.4
53	14.42	10721	329	400	380.2	110	23.68	7062	264	522	428.2
54	14.69	6420	268	426	379.0	111	24.36	10831	400	472	446.8
55	14.69	4923	209	381	354.4	112	27.39	11381	422	413	384.2
56	14.72	6597	245	489	460.0	113	28.06	10253	425	413	385.8
57	14.75	10886	350	413	395.6						

Solution 12.4

A plot of dry matter against the maximum stem length (Figure S12.4.1) shows a noisy relationship, which appears approximately linear, albeit with some suggestion of curvature. It is clear that an SLR with maximum stem length as the explanatory variate is not going to give a wholly accurate prediction of total dry matter, but we will nevertheless fit the model to evaluate its performance.

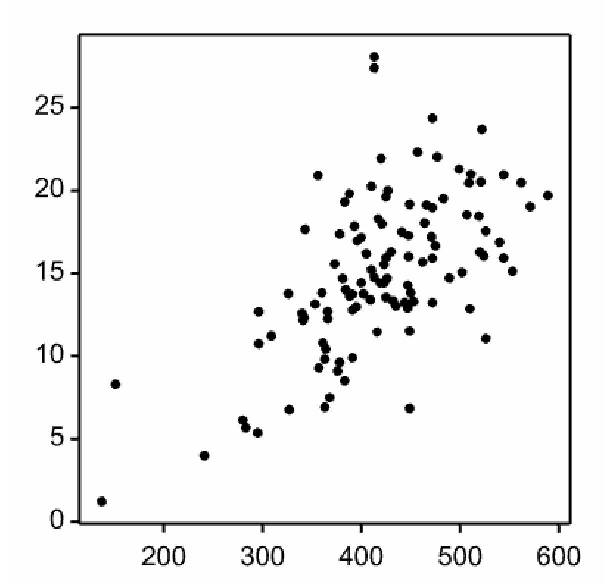

Figure S12.4.1. Dry matter plotted against maximum stem length

A SLR for dry matter in terms of maximum stem length takes the symbolic form

Response: *DryMatter*
Explanatory component: *[1] + MaxLength*

and mathematical form

$$DryMatter_i = \alpha + \beta MaxLength_i + e_i$$

for i = 1…113, corresponding to the 113 leaves in the sample, using obvious variable names and the notation of Chapter 12. The summary ANOVA table for this model is Table S12.4.1. There is strong evidence of a linear relationship between the two variables ($F_{1,111}$ = 64.77 with $P < 0.001$). However, although this evidence is very strong, the percentage variance accounted for by the model is relatively low (adjusted R^2 = 0.363), reflecting the scatter seen in Figure S12.4.1.

Table S12.4.1 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	919.144	919.144	64.768	< 0.001
Residual	111	1575.230	14.191		
Total	112	2494.373			

Table S12.4.2 Parameter estimates with standard errors (SE), t-statistics (t) and observed significance levels (*P*) for a SLR model for total biomass (variate *DryMatter*) with explanatory variate *MaxLength*.

Term	Parameter	Estimate	SE	t	*P*
[1]	α	-0.36	1.95	-0.19	0.853
MaxLength	β	0.03665	0.00455	8.05	< 0.001

The parameter estimates are in Table S12.4.2. The estimate of the intercept (α) is negative, but this corresponds to an extrapolation far outside the range of the explanatory variate (137 to 589) and so we are not concerned. The estimate of the slope (β) is positive indicating larger biomass for trees with longer longest stems.

A composite set of residual plots for this SLR model are shown in Figure S12.4.2. The fitted value plot suggests a small degree of curvature in the relationship, but this is not strong. There are two large residuals with a fitted value close to 15 which suggest there are some samples that do not correspond well to the predicted values from this model.

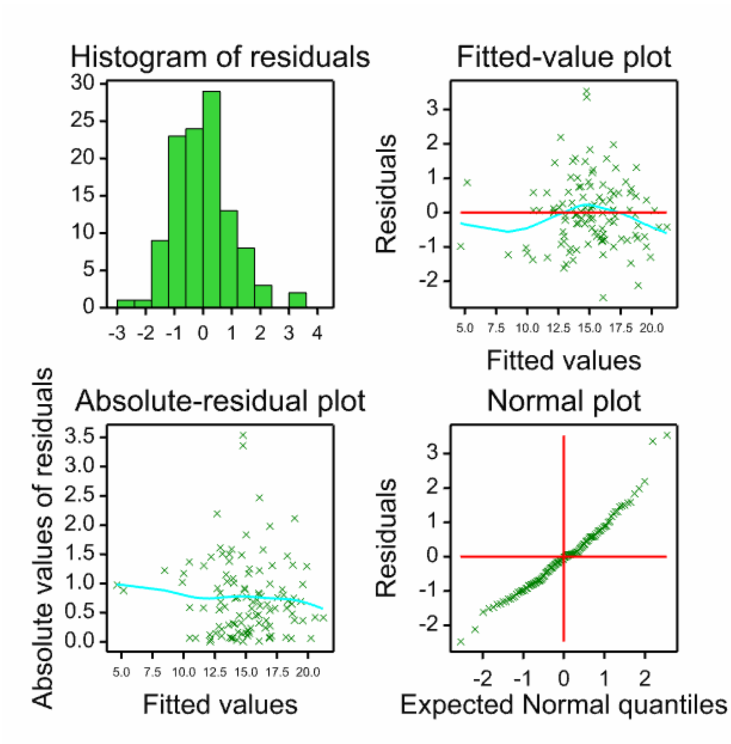

Figure S12.4.2 Composite set of residual plots for SLR with dry matter as the response and maximum stem length as the explanatory variate.

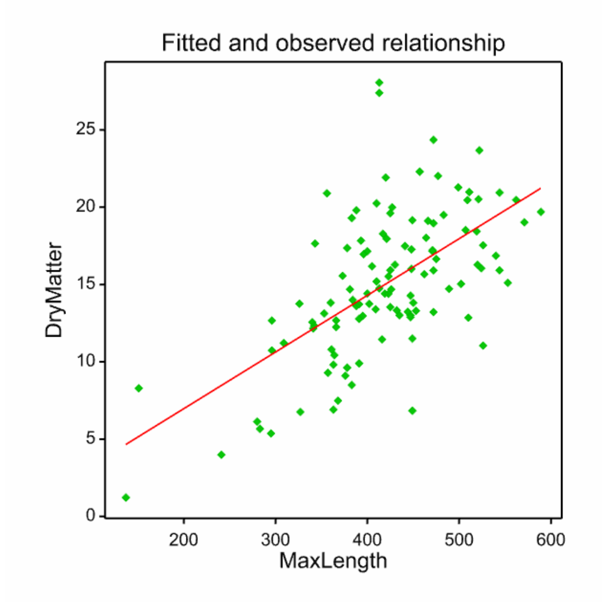

Figure S12.4.3. Observed dry matter with fitted SLR using maximum stem length as the explanatory variate.

The fitted model is shown with the data in Figure S12.4.3. The two observations with dry matter values > 25 and maximum length ~400 correspond to the two outliers in the residual plots (Figure S12.4.2). A predictive model for dry matter from this SLR can be written as

$$\hat{\mu}(MaxLength) = -0.36 + 0.03665 MaxLength$$

However, given the low percentage variance accounted for, the outlying observations and the availability of other potential surrogate variables, we would probably not accept maximum stem length as a surrogate variable. Instead, we would investigate the other variables to see if any of them, or a combination of them, would do better, and then weigh any potential gain in accuracy against the cost of making those measurements. We investigate these other models in Exercises 13.5 and 14.5.