Solution to Exercise 18.11 (Version 1, 31/8/15)

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 18.11 (Data: courtesy D. Gray, Horticulture Research International)

The viability of carrot seed depends greatly on the conditions under which it is stored. Replicate batches of 100 seeds were stored in each of four different conditions (labelled A–D). Four replicate batches were sampled from each condition at pre-specified times: conditions A and B were sampled approximately every 60 days and conditions C and D were sampled approximately every 30 days, and the number of non-viable seeds was evaluated. File CARROT.DAT contains unit numbers (*ID*), the structural factors (Batch, Sample), explanatory variables (factor Condition, variate *Days*) and response (variate *Count*). Use a GLM to model the number of non-viable seeds over time in each condition and check the fit of the model carefully. Is there any evidence of model misspecification? Identify any features of the data that are incompatible with the GLM.

Data 18.11 (CARROT.DAT) Counts (column Ct) of non-viable carrot seed in replicate batches (column B) evaluated at different times (measured in days, column D or by sample number, column S) following storage under conditions A-D (column C).

ID	B	S	C	D	Ct
1	1	1	A	0	12
2	1	2	A	61	17
3	1	3	A	119	14
4	1	4	A	181	12
5	1	5	A	242	25
6	1	6	A	306	23
7	1	7	A	364	41
8	1	8	A	424	41
9	1	9	A	515	80
10	1	10	A	608	93
11	1	11	A	699	97
12	1	12	A	790	100
13	1	13	A	885	100
14	1	14	A	976	100
15	1	15	A	1064	100
16	2	1	A	0	10
17	2	2	A	61	21
18	2	3	A	119	18
19	2	4	A	181	21
20	2	5	A	242	24
21	2	6	A	306	35
22	2	7	A	364	33
23	2	8	A	424	43
81	2	6	B	306	17
82	2	7	B	364	17
83	2	8	B	424	24
84	2	9	B	515	28
85	2	10	B	608	39
86	2	11	B	699	50
87	2	12	B	790	68
88	2	13	B	885	60
89	2	14	B	976	90
90	2	15	B	1064	97
91	3	1	B	0	7
92	3	2	B	61	10
93	3	3	B	119	10
94	3	4	B	181	17
95	3	5	B	242	7
96	3	6	B	306	11
97	3	7	B	364	17
98	3	8	B	424	21
99	3	9	B	515	27
100	3	10	B	608	36
101	3	11	B	699	50
102	3	12	B	790	66
103	3	13	B	885	70
161	3	9	C	242	42
162	3	10	C	273	57
163	3	11	C	306	76
164	3	12	C	334	76
165	3	13	C	364	93
166	3	14	C	391	93
167	3	15	C	424	100
168	3	16	C	453	100
169	4	1	C	0	19
170	4	2	C	28	14
171	4	3	C	61	14
172	4	4	C	89	11
173	4	5	C	119	22
174	4	6	C	151	22
175	4	7	C	181	34
176	4	8	C	213	32
177	4	9	C	242	49
178	4	10	C	273	59
179	4	11	C	306	70
180	4	12	C	334	88
181	4	13	C	364	96
182	4	14	C	391	96
183	4	15	C	424	100

ID	B	S	C	D	Ct
24	2	9	A	515	77
25	2	10	A	608	94
26	2	11	A	699	99
27	2	12	A	790	100
28	2	13	A	885	100
29	2	14	A	976	100
30	2	15	A	1064	100
31	3	1	A	0	9
32	3	2	A	61	8
33	3	3	A	119	22
34	3	4	A	181	19
35	3	5	A	242	26
36	3	6	A	306	26
37	3	7	A	364	37
38	3	8	A	424	50
39	3	9	A	515	76
40	3	10	A	608	95
41	3	11	A	699	99
42	3	12	A	790	98
43	3	13	A	885	100
44	3	14	A	976	100
45	3	15	A	1064	100
46	4	1	A	0	19
47	4	2	A	61	16
48	4	3	A	119	11
49	4	4	A	181	12
50	4	5	A	242	27
51	4	6	A	306	32
52	4	7	A	364	40
53	4	8	A	424	38
54	4	9	A	515	81
55	4	10	A	608	86
56	4	11	A	699	99
57	4	12	A	790	100
58	4	13	A	885	100
59	4	14	A	976	100
60	4	15	A	1064	100
61	1	1	B	0	11
62	1	2	B	61	11
63	1	3	B	119	18
64	1	4	B	181	9
65	1	5	B	242	15
66	1	6	B	306	16
67	1	7	B	364	19
68	1	8	B	424	26
69	1	9	B	515	22
70	1	10	B	608	35
71	1	11	B	699	49
72	1	12	B	790	49

ID	B	S	C	D	Ct
104	3	14	B	976	91
105	3	15	B	1064	96
106	4	1	B	0	11
107	4	2	B	61	4
108	4	3	B	119	8
109	4	4	B	181	5
110	4	5	B	242	20
111	4	6	B	306	11
112	4	7	B	364	17
113	4	8	B	424	17
114	4	9	B	515	22
115	4	10	B	608	38
116	4	11	B	699	47
117	4	12	B	790	58
118	4	13	B	885	69
119	4	14	B	976	85
120	4	15	B	1064	98
121	1	1	C	0	12
122	1	2	C	28	8
123	1	3	C	61	15
124	1	4	C	89	15
125	1	5	C	119	20
126	1	6	C	151	26
127	1	7	C	181	31
128	1	8	C	213	37
129	1	9	C	242	44
130	1	10	C	273	48
131	1	11	C	306	75
132	1	12	C	334	82
133	1	13	C	364	98
134	1	14	C	391	94
135	1	15	C	424	100
136	1	16	C	453	100
137	2	1	C	0	10
138	2	2	C	28	10
139	2	3	C	61	11
140	2	4	C	89	10
141	2	5	C	119	13
142	2	6	C	151	26
143	2	7	C	181	21
144	2	8	C	213	33
145	2	9	C	242	47
146	2	10	C	273	63
147	2	11	C	306	77
148	2	12	C	334	85
149	2	13	C	364	96
150	2	14	C	391	96
151	2	15	C	424	100
152	2	16	C	453	100

ID	B	S	C	D	Ct
184	4	16	C	453	100
185	1	1	D	0	11
186	1	2	D	28	14
187	1	3	D	61	9
188	1	4	D	89	14
189	1	5	D	119	17
190	1	6	D	151	22
191	1	7	D	181	32
192	1	8	D	213	57
193	1	9	D	242	67
194	1	10	D	273	71
195	1	11	D	306	88
196	1	12	D	334	97
197	1	13	D	364	99
198	1	14	D	391	100
199	2	1	D	0	9
200	2	2	D	28	8
201	2	3	D	61	8
202	2	4	D	89	17
203	2	5	D	119	17
204	2	6	D	151	28
205	2	7	D	181	47
206	2	8	D	213	57
207	2	9	D	242	81
208	2	10	D	273	81
209	2	11	D	306	90
210	2	12	D	334	99
211	2	13	D	364	99
212	2	14	D	391	100
213	3	1	D	0	7
214	3	2	D	28	8
215	3	3	D	61	10
216	3	4	D	89	12
217	3	5	D	119	18
218	3	6	D	151	24
219	3	7	D	181	40
220	3	8	D	213	62
221	3	9	D	242	71
222	3	10	D	273	89
223	3	11	D	306	97
224	3	12	D	334	98
225	3	13	D	364	99
226	3	14	D	391	100
227	4	1	D	0	11
228	4	2	D	28	6
229	4	3	D	61	13
230	4	4	D	89	9
231	4	5	D	119	17
232	4	6	D	151	25

ID	B	S	C	D	Ct	ID	B	S	C	D	Ct	ID	B	S	C	D	Ct
73	1	13	B	885	56	153	3	1	C	0	9	233	4	7	D	181	42
74	1	14	B	976	80	154	3	2	C	28	15	234	4	8	D	213	62
75	1	15	B	1064	98	155	3	3	C	61	9	235	4	9	D	242	68
76	2	1	B	0	9	156	3	4	C	89	13	236	4	10	D	273	79
77	2	2	B	61	7	157	3	5	C	119	20	237	4	11	D	306	86
78	2	3	B	119	8	158	3	6	C	151	17	238	4	12	D	334	97
79	2	4	B	181	14	159	3	7	C	181	19	239	4	13	D	364	94
80	2	5	B	242	16	160	3	8	C	213	37	240	4	14	D	391	100

Solution 18.11

Within each storage condition, the design is effectively CRD with multiple replicates (batches) chosen at random to be evaluated at each sampling point. We therefore ignore the sample and batch factors in the data file for the purposes of analysis. The counts are the number of non-viable seeds out of 100 in each batch, so we assume a Binomial distribution. We want to model the proportion of non-viable seeds as a function of time (days in storage) which may depend on storage conditions, so we write our initial model as

Response variable:	*Count*
Probability distribution:	Binomial (number of tests = 100)
Link function:	logit
Explanatory component:	[1] + *Days**Condition

This model allows separate lines on the logit scale for the proportion of non-viable seeds. The data is shown in Figure S18.11.1, and shows a logit-shape response to time of storage and that viability appears to differ greatly between storage conditions.

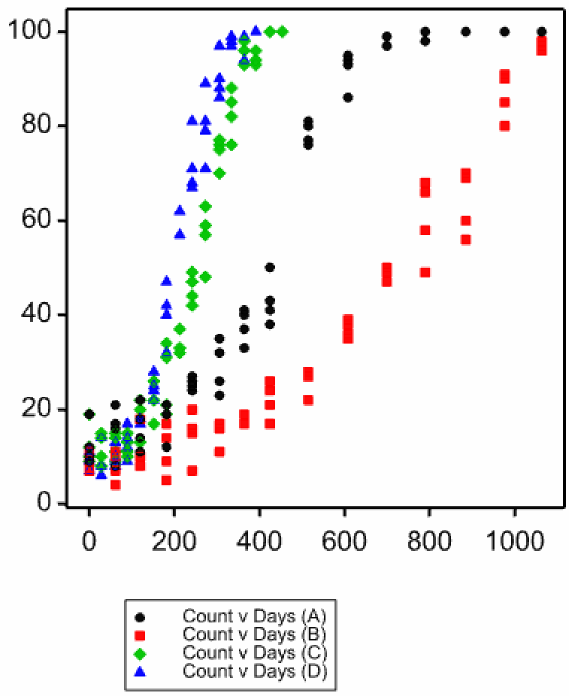

Figure S18.11.1 Number of non-viable seeds per batch plotted against number of days in storage.

We start by fitting this separate lines model and check for over-dispersion. The residual deviance from this model is 948.8 with 232 df, and when compared to a chi-square distribution with 232 df gives strong evidence of over-dispersion ($P < 0.001$). We refit the model with over-dispersion estimated to obtain the sequential ANODEV table shown in Table S18.11.1. There is very strong evidence that separate slopes and intercepts are required. A composite set of residual plots is in Figure S18.11.2. There is a clear trend in the fitted value plot that we suspect is due to lack-of-fit of the separate lines model to the observed data. We investigate this further by inspecting plots of residuals against days within each treatment group (Figure S18.11.3) and of the data with the fitted model (Figure S18.11.4).

Table S18.11.1 A sequential ANODEV table for separate lines GLM for proportion of non-viable seeds with Binomial distribution and logit link.

Source of variation	df	Deviance	Mean deviance	Deviance Ratio	*P* (F prob.)
+ *Days*	1	6282.412	6828.412	1669.67	< 0.001
+ Condition	3	4735.465	1578.488	385.97	< 0.001
+ *Days*.Condition	3	1897.916	632.639	154.69	< 0.001
Residual	232	948.805	4.090		
Total	239	14410.599			

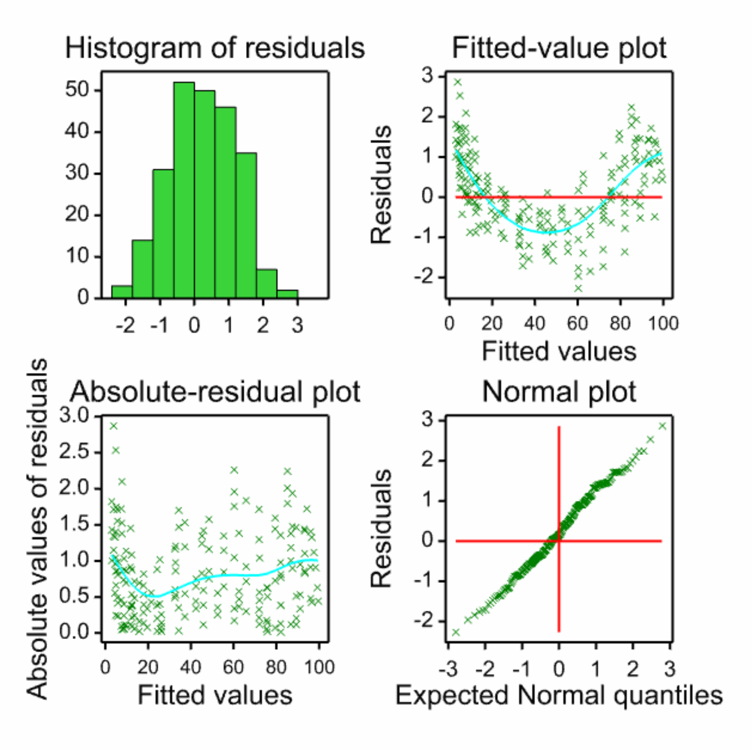

Figure S18.11.2 Composite set of residual plots based on standardized deviance residuals for GLM for separate lines model for proportion of non-viable seeds with Binomial distribution and logit link.

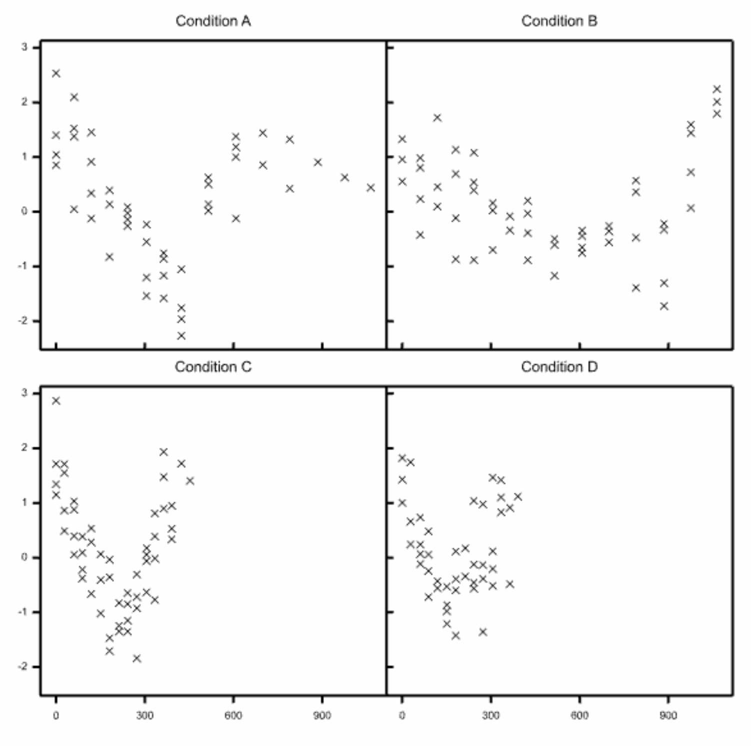

Figure S18.11.3 Standardized deviance residuals from separate lines model plotted against days.

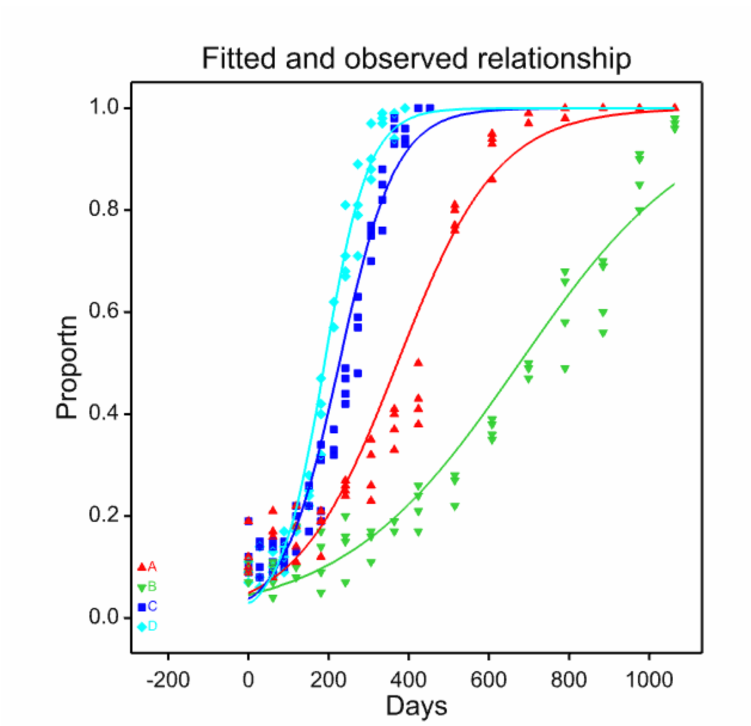

Figure S18.11.4 Fitted separate lines model with observed data for each condition.

Figure S18.11.3 shows strong curvature in the pattern of the residuals over time which reflects lack of fit to linear trend on the logit scale. The impact on the fitted model can be seen in Figure S18.11.4.

Lack of fit at the start of the experiment is particularly clear, as the observed patterns tail off above zero; some seeds are not viable even before storage. This phenomenon is sometimes called ‘control mortality’, ie. death in the absence of any treatment, and often causes problems with simple logistic regression models when it is present. It is possible to extend the model to include this parameter, but not within the framework of GLMs (this requires a non-linear model with a Binomial distribution). Instead, we will notice the quadratic shape of the residual plots in Figure S18.11.3, and investigate whether adding quadratic terms into the model can improve its fit (see also Section 17.1.2). If we calculate a new variate as

$$DaySqrd = Day * Day$$

then we can write this new model in symbolic form as

Response variable:	*Count*
Probability distribution:	Binomial (number of tests = 100)
Link function:	logit
Explanatory component:	*[1]* + Condition*(*Days*+*DaySqrd*)

A sequential ANODEV table for this model is in Table S18.11.2. The residual mean deviance is much smaller, but a test shows there is still evidence of over-dispersion. Residual plots from this model are shown in Figure S18.11.5 and the fitted model is shown in Figure S18.11.6. In both cases, it is clear that the fit is greatly improved. We can also test formally for lack of fit to these fitted quadratic curves using a factor version of the *Days* variate (called fDay). The explanatory component of the model then becomes

Explanatory component: *[1]* + Condition*(*Days*+*DaySqrd*+fDay)

The lack of fit terms, fDay and Condition.fDay, then give strong evidence of lack of fit but the mean deviance associated with both terms is relatively small, and the differences in fitted values from the quadratic models are small. We might reasonably accept the quadratic model as an adequate description of the patterns whilst recognising that the fit is not perfect.

Table S18.11.2 A sequential ANODEV table for GLM with quadratic terms for proportion of non-viable seeds with Binomial distribution and logit link.

Source of variation	df	Deviance	Mean deviance	Deviance Ratio	*P* (F prob.)
+ Condition	3	666.912	222.304	153.04	< 0.001
+ *Days*	1	10896.965	10896.965	7501.66	< 0.001
+ *DaySqrd*	1	383.495	383.495	264.00	< 0.001
+ Condition.*Days*	3	1817.836	605.945	417.14	< 0.001
+ Condition.*DaySqrd*	3	314.195	104.732	72.10	< 0.001
Residual	228	331.195	1.453		
Total	239	14410.599			

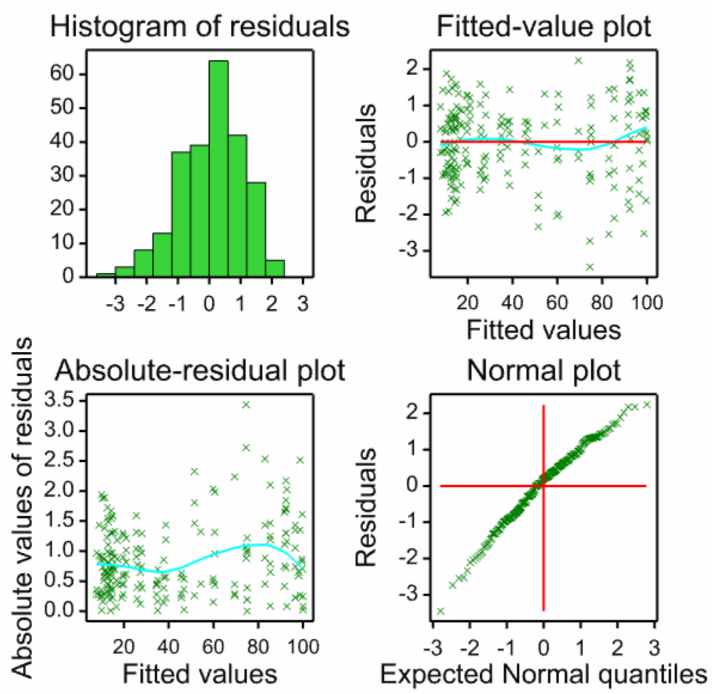

Figure S18.11.5 Composite set of residual plots based on standardized deviance residuals for GLM for separate quadratic curves model for proportion of non-viable seeds with Binomial distribution and logit link.

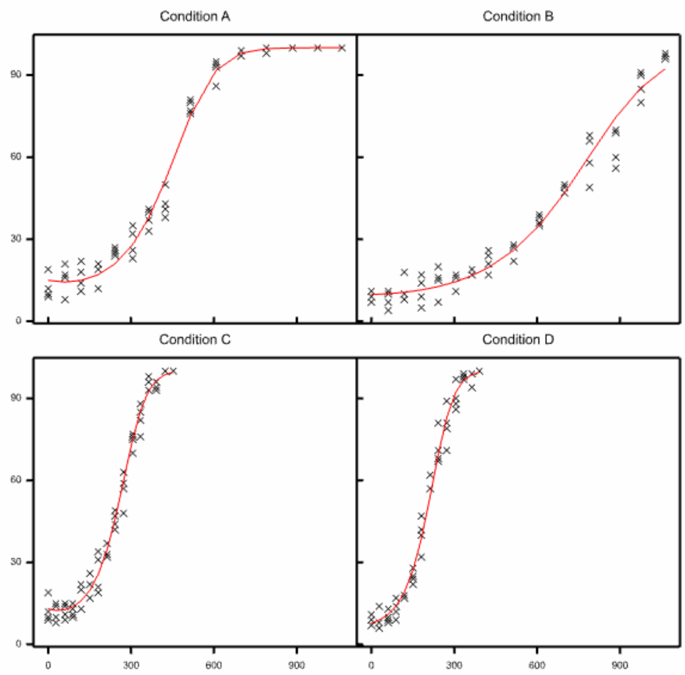

Figure S18.11.6 Fitted separate quadratic curves model with observed data.