Solution to Exercise 18.3 (Version 1, 21/8/15)

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 18.3*

A pilot study investigated the period of leaf wetness required to successfully infect leaves with a foliar disease. Trays of four young plants with four leaves were sprayed with inoculum and then kept wet for a period of 16, 24, 48 or 72 hours. The experiment used a CE cabinet with four shelves and was designed as a RCBD, with shelves used as blocks. File WETNESS.DAT holds the unit numbers (*ID*), structural factors (Shelf, Tray) with the wetness period (variate *Wetness*) and number of leaves infected (variate *NInf*, number out of 16). What distribution might you expect the number of infected leaves to follow? Use a suitable GLM to model the number of infected leaves in each tray, taking account of the design structure by including shelves in the model. Check for evidence of over-dispersion, check residual plots and carry out a formal test for lack of fit. Is there any evidence that wetness period affects the number of infected leaves? Predict the probability that a leaf is infected after 36 hours of wetness, and give confidence limits for this prediction.

Data 18.3 (WETNESS.DAT)

ID	Shelf	Tray	Wetness	NInf	ID	Shelf	Tray	Wetness	NInf
1	1	1	72	14	9	3	1	72	13
2	1	2	48	12	10	3	2	48	10
3	1	3	24	8	11	3	3	24	7
4	1	4	16	2	12	3	4	16	6
5	2	1	72	16	13	4	1	72	6
6	2	2	48	11	14	4	2	48	3
7	2	3	24	8	15	4	3	24	0
8	2	4	16	3	16	4	4	16	0

Solution 18.3

Note: the question omits the detail that the number of infected leaves for each treatment was assessed at the end of the experiment, after 10 days.

There are 16 leaves in each tray (four per plant) and each is assessed as healthy or infected. If we assume that the 16 leaves in each tray develop disease independently, then the number of infected leaves might be expected to follow a binomial distribution, with the hypothesis that probability of infection is related to wetness period. We can therefore use a GLM to model the number of infected leaves per tray. Figure S18.3.1 shows the proportion of infected leaves plotted against wetness period with the shelves indicated by different colours/symbols in the left-hand plot, with logit(proportion of infected leaves) in the right-hand plot. Shelf 4 (blue triangles) has a lower proportion of infected leaves than the other shelves, and there is an approximately linear increase in the proportion infected which is possibly slightly improved by the logit transformation.

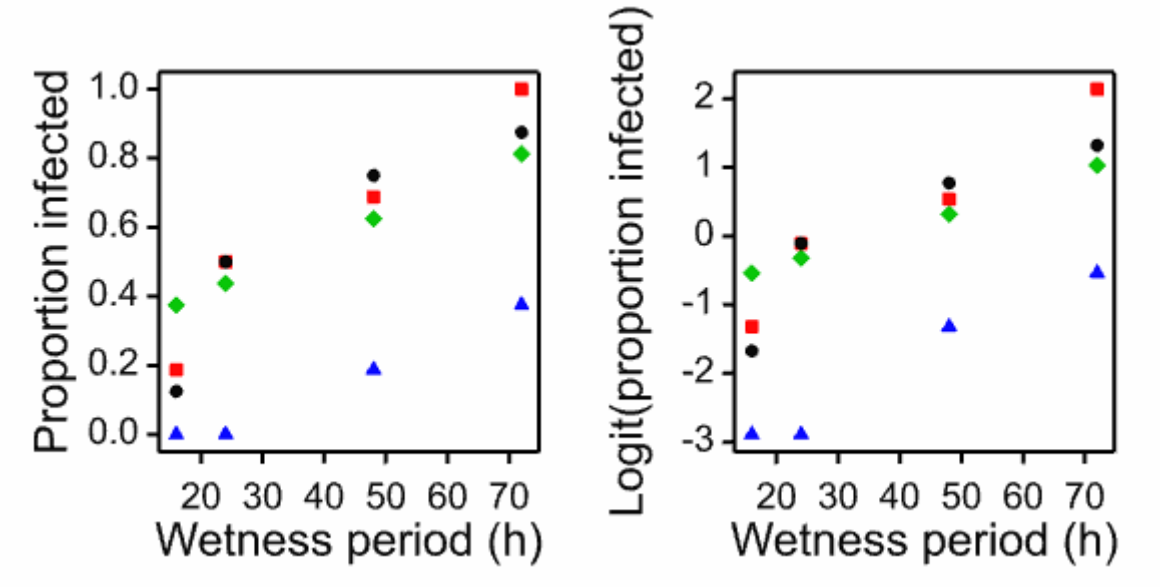

Figure S18.3.1 Proportion of infected leaves and logit(proportion infected leaves) plotted against wetness period and coloured by shelf number.

These plots suggest that a binomial GLM with logit link may well give a good description of the data, and illustrate the importance of accounting for the blocking (factor Shelf) as a source of extra variation.

As there are replicate observations for each wetness period, there are two models we might consider. Our first option is to fit means for each wetness period, ie. using a factor to represent the different wetness periods. This model is sensible only when we have replicate observations within each group. The second option is to fit a linear relationship (on the logit scale), ie. using a variate holding the numeric values of each wetness period. This second option uses fewer parameters and has less flexibility, and will give a poor fit if the relationship on the logit scale is not linear. We will use the first option, which has no issues with possible lack-of-fit, to assess whether over-dispersion is present. After defining factor fWetness to have a separate level for each wetness period, we can write this model in symbolic form as

Response variable:	*NInf*
Probability distribution:	Binomial (Number of tests = 16)
Link function:	logit
Explanatory component:	*[1]* + Shelf + fWetness

The ANODEV table for this model is Table S18.3.1. The residual mean deviance is 10.93 with 9 df and gives no indication of over-dispersion when compared with a chi-squared distribution with 9 df (P = 0.281). A composite set of residual plots from this model are shown in Figure S18.3.2 and although they are far from perfect, we accept that this is a small data set and that there is no obvious cause for concern. We can therefore move forward to investigate the relationship between proportion of infected leaves and wetness period, but we can extract further information from this model to inform the process. We first check the sequential ANOVA table from our current model (Table S18.3.2) to see if there is any evidence from this analysis of differences in infection rates related to wetness periods. It is important here to have fitted the shelf factor first, so we can examine the effects of wetness period after accounting for block (shelf) differences.

Table S18.3.1 Summary ANODEV table for GLM with Binomial distribution and logit link for number of infected leaves with explanatory factors Shelf and fWetness, assuming no over-dispersion.

Source of variation	df	Deviance	Mean deviance	*P* (Chi-squared prob.)
Model	6	104.59	17.431	< 0.001
Residual	9	10.93	1.214	
Total	15	115.51		

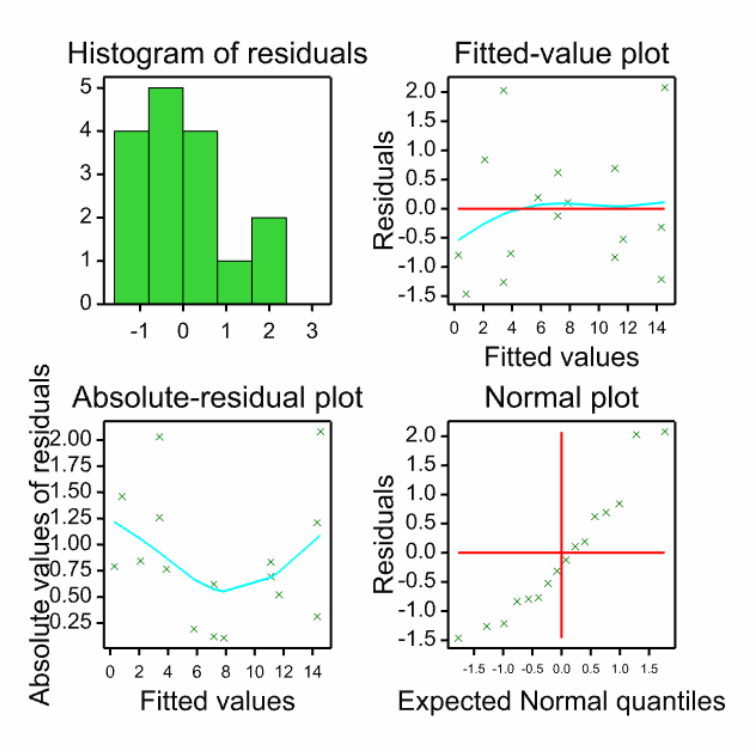

Figure 18.3.2. Composite set of residual plots from GLM with binomial distribution and logit link for number of infected leaves, fitting means for blocks (shelves) and treatments (wetness periods).

Table S18.3.2 Sequential ANODEV table for GLM with Binomial distribution and logit link for number of infected leaves with explanatory factors Shelf and fWetness, assuming no over-dispersion.

Source of variation	df	Deviance	Mean deviance	*P* (Chi-squared prob.)
+ Shelf	3	39.744	13.224	< 0.001
+ fWetness	3	64.842	21.614	< 0.001
Residual	9	10.926	1.214	
Total	15	115.513		

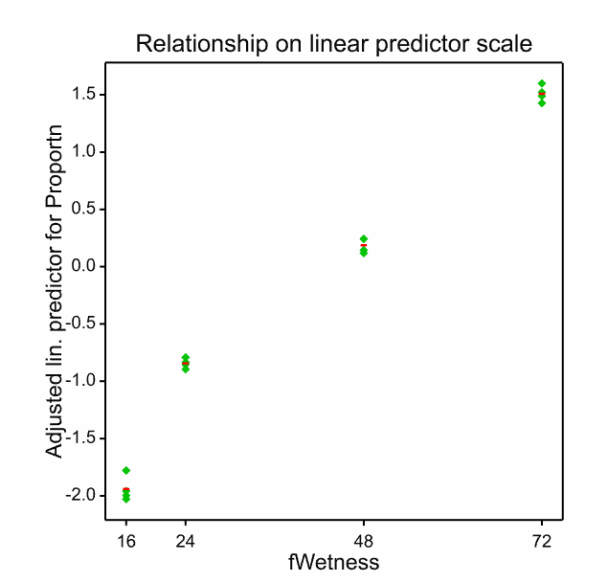

Figure S18.3.3 Predicted population means for wetness periods on logit scale.

The sequential ANODEV table shows strong evidence of differences in infection rate among wetness periods (mean deviance = 21.614 with 3 df, $P < 0.001$ when compared to a chi-square distribution with 3 df). A plot of the predicted population means (averaged over shelves, Figure S18.3.3) on the logit scale shows a trend that is close to linear, so we will investigate next whether a linear relationship gives an adequate description. We will do this by fitting a linear relationship (on the logit scale) with wetness period by using the Wetness variate in the model, and will also include the fWetness factor to test for lack-of-fit. Our new model can be written in symbolic form as

Response variable:	*NInf*
Probability distribution:	Binomial (Number of tests = 16)
Link function:	logit
Explanatory component:	*[1]* + Shelf + *Wetness* + fWetness

The ANODEV table from this model is Table S18.3.3. This table has partitioned the deviance accounted for by wetness periods into that associated with linear trend (mean dev = 62.604, $P < 0.001$) and a remainder, representing variation around that linear trend (mean dev = 1.119, $P = 0.327$). We can therefore conclude that there is no evidence of lack of fit for the linear relationship, and proceed to fit that model.

Table S18.3.3 Sequential ANODEV table for GLM with Binomial distribution and logit link for number of infected leaves with explanatory factor Shelf, variate *Wetness* and factor fWetness to test for lack of fit, assuming no over-dispersion.

Source of variation	df	Deviance	Mean deviance	*P* (Chi-squared prob.)
+ Shelf	3	39.744	13.224	< 0.001
+ *Wetness*	1	62.604	62.604	< 0.001
+ fWetness	2	2.238	1.119	0.327
Residual	9	10.926	1.214	
Total	15	115.513		

The linear relationship model is written in symbolic form as

Response variable:	*NInf*
Probability distribution:	Binomial (Number of tests = 16)
Link function:	logit
Explanatory component:	*[1]* + Shelf + *Wetness*

The ANODEV table for this model is Table S18.3.4 and the estimated parameters are in Table S18.3.5. We can write this model for the data in mathematical form as

$$NInf_{ij} \sim \text{Binomial}(16, p_{ij}); \ \eta_{ij} = \text{logit}(p_{ij}) = \alpha_1 + \nu_i + \beta x_{ij}.$$

where

- $NInf_{ij}$ is the number of infected leaves in the jth tray on the ith shelf
- p_{ij} is the expected proportion of infected leaves in the jth tray on the ith shelf
- η_{ij} is the logit transform of the expected proportion p_{ij}
- α_1 is the intercept for shelf 1
- ν_i is the difference in intercept for the ith shelf (i=1…4) compared to the first shelf (ν_1=0)
- β is the estimated slope (on the logit scale) of the relationship with wetness period
- x_{ij} is the wetness period used for the jth tray on the ith shelf

These parameter labels match those in Table S18.3.5. For an extra 10 hours of wetness, we therefore expect an increase of 0.57 units on the logit scale. The fitted model is shown on the logit and proportion scales in Figure S18.3.4 and appears to give a reasonable fit to the trend.

Table S18.3.4 Sequential ANODEV table for GLM with Binomial distribution and logit link for number of infected leaves with explanatory factor Shelf and variate *Wetness*, assuming no over-dispersion.

Source of variation	df	Deviance	Mean deviance	*P* (Chi-squared prob.)
+ Shelf	3	39.744	13.224	< 0.001
+ *Wetness*	1	62.604	62.604	< 0.001
Residual	11	13.165	1.197	
Total	15	115.513		

Table S18.3.5 Parameter estimates for GLM with Binomial distribution and logit link for number of infected leaves with explanatory factor Shelf and variate *Wetness*, using first-level-zero parameterization.

Term	Parameter	Estimate	SE	t	*P*
[1]	α_1	-1.88	0.410	-4.576	< 0.001
Shelf 2	ν_2	0.17	0.414	0.413	0.679
Shelf 3	ν_3	0.00	0.413	0.00	1.000
Shelf 4	ν_4	-2.74	0.526	-5.207	< 0.001
Wetness	β	0.0567	0.00834	6.806	< 0.001

We can make predictions for wetness period for a "typical shelf" by averaging over shelves before back-transformation (see Section 18.2.5). Our predictive model therefore becomes

$$\hat{\eta}(x) = \hat{\alpha}_1 + \frac{1}{4}\sum_{i=1}^{4}\hat{v}_i + \hat{\beta}x = -2.519 + 0.0567\,x\,.$$

Predicting the logit proportion of infected leaves at 36 hours gives

$$\hat{\eta}(x = 36) = -2.519 + 0.0567 \times 36 = -0.4762$$

With the aid of statistical software, we can find the SE for this prediction as SE = 0.1636 and get a 95% confidence interval for this value as (-0.8362, -0.1162) in the usual way (using the 97.5th percentile of the t distribution with 11 df). We can back-transform onto the original scale (via the inverse logit transformation, see Equation 18.2) to get our estimated proportion of infected leaves for 36 hours wetness as 0.383, with 95% confidence interval (0.302, 0.471).

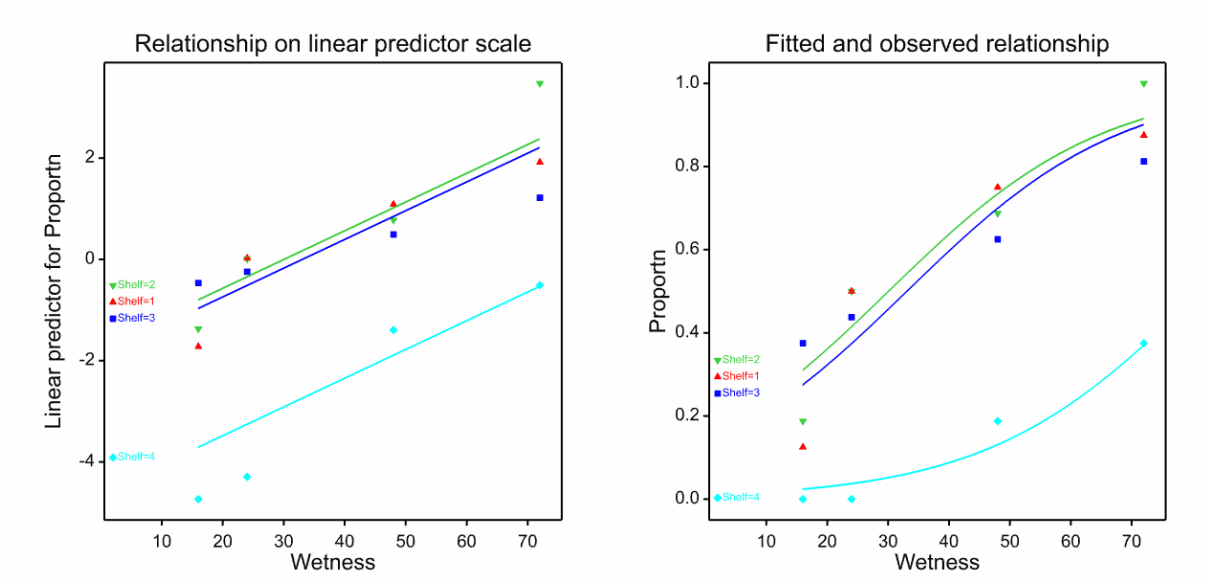

Figure S18.3.4 Fitted model with separate intercepts for blocks and linear relationship (on logit scale) with wetness period on logit scale (left) and proportion scale (right). The fitted models for shelves 1 and 3 (red and blue) are coincident.