- Comparisons between levels of `tint` or `target` are made several times for each of the 26 individuals, and are relatively consistent from one individual to another. Standard errors for these comparisons are small – in the range 0.042 - 0.058.

Statistical variation cannot be convincingly ruled out as the explanation for the effects that stand out most strongly in the graphs. The graphs are not designed to highlight the consistency with which individuals respond to differences between levels of tinting and target contrast.

7.6 A mixed model with a betabinomial error

7.6.1 The betabinomial distribution

Data that has the form required for analysis as binomial (e.g., x insects dead out of n) will commonly not follow a binomial distribution. An immediate indication of this is a variance that is larger than the variance for a binomial distribution with the same mean. The data is then said to be over-dispersed relative to the binomial. Subsection 5.1.3 noted the possibility of using the beta-binomial distribution to model data that over-dispersed relative to the binomial. The total number of "successes" out of n is modeled as the sum from n Bernoulli trials (which have a 0/1 outcome), with the probability varying from one trial to another according to a beta distribution.

The function `glmmTMB`, as implemented in the `glmmTMB` package, allows the modeling of the scale parameter as a function of explanatory variables. This will be important for the insect mortality data considered here, and for other datasets for data of this type where, at high mortalities, the overdispersion factor is high at midrange mortalities and close to 1 (i.e., close to what would be expected for a binomial distribution) at high mortalities.

For more details on the beta-binomial, see for example Morgan (1992). Morgan and Ridout (2008) is interesting because it compares use of a binomial distribution, a beta-binomial, and a mixture of the two distributions.

In the sequel, it will be necessary to have the `qra` (quantal response analysis), `glmmTMB`, and `bblme` packages installed. The `glmmTMB` and `bblme` packages can be installed from CRAN. At this time, `qra` must be installed from Gitlab, thus:

```
devtools::install_git("https://gitlab.com/daagur/qra.git")
```

Notation

For present purposes, it is the scale parameter that is of interest. Take the location parameter π, with observed value P, to be the expected mortality. Then, in notation consistent with that used in `glmmTMB`:

$$\text{var}[P] = \frac{\pi(1-\pi)}{n}(1+(n-1)\rho), \text{ where } \rho = (1+e^{\eta})^{-1}$$

The parameter ρ is the intra-class correlation. It can be used, as an alternative to η, as the scale parameter. The function `glmmTMB()` has provision to model variation in η, and hence ρ, as a function of one or more explanatory variables and/or factors.

The overdispersion factor is $\phi = 1+(n-1)\rho$. An important difference from the dispersion as defined for quasibinomial models (see `?quasibinomial`) is that it is now a function of

n, rather than constant as n varies.

The variance can never be less than $\pi(1-\pi)\rho$. The fraction by which it is increased above this minimum is $(1-\rho)n^{-1}$. If $\rho > 0$, then as the sample size n increases, there comes a point at which any further increase in n gives only a very slight further increase in the variance. It may be that the betabinomial model is in this respect too pessimistic. In the present state of the are, packaged code appears not to offer good alternatives, such as perhaps the mixture of binomial and betabinomial that is described in Morgan and Ridout (2008).

Source of data

Disinfestion is the removal or disabling of insect pests or pathogens from produce, allowing it to be transported across international boundaries. Cold storage, usually applied while the produce is in transit, is now a common and effective approach that avoids the use of chemical fumigants.

The dataset `qra::HawCon` is from experimental work that was designed to assess the lengths of times in cold storage that, for the species/lifestage combinations investigated, might be effective. We thank Dr Peter Follett for allowing us to use this dataset. Data is for four life-stages of each of two species, thus:

- Species are: Mediterranean fruit fly (MedFly); and Melon fly (MelonFly)
- Life-stages are: Egg, L1 (Larval 1), L2, L3
  - There is one replicate of L1, 3 replicates of others
- Mortality is for varying times in cool-storage at 1.5 - 2^{o}C

Code that provides the `HawCon` data, in the form required, is:

```
HawCon <- within(as.data.frame(qra::HawCon), {
  trtGp <- paste0(CN, LifestageTrt)
  trtGp <- gsub("Fly","",trtGp)
  trtGp <- factor(trtGp, levels=sort(unique(trtGp))[c(1,5,2,6,3,7,4,8)])
  gp <- paste0(CN, LifestageTrt,":",RepNumber)
  scTime <- scale(TrtTime)
  ## Model may fit more readily with a centered and scaled variable
})
```

Fit `glmmTMB` model – fruitfly data

```
## Load packages that will be used
suppressMessages(library(lme4))
suppressMessages(library(glmmTMB))
suppressMessages(library(DHARMa))
```

Choice of model

For both the complementary log-log link and logit links, we now try the alternatives:

- Separate lines for each treatment group;
- Separate lines for each treatment group, plus a 2-degree polynomial function of time in cold storage that applies a common adjustment across all treatment groups. (For this "added curve" model, the suffix "2s" is added to the model name.)

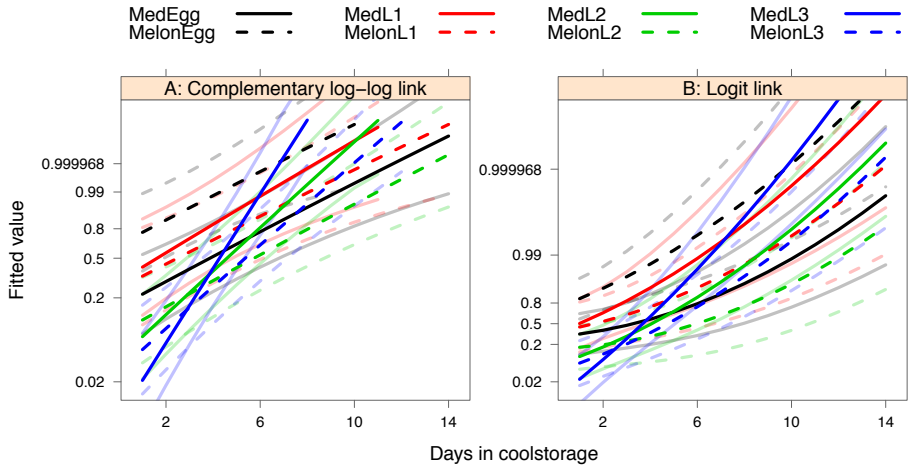

Figure 7.9: Fitted quadratic curves, for the model where the curve shifts up/down a/c variety. Panel A is for the model that uses a complementary log-log (cloglog) link, while Panel B is for a logit link.

The scale parameter is in each case modeled as a 2-degree polynomial function of time in coolstorage, with a different multiplier for each different treatment group. (Effects that are additive on the logarithmic link scale become multipliers when unlogged.)

```
ablin.TMBclog <- glmmTMB(cbind(Dead,Live)~0+trtGp/TrtTime+(TrtTime|gp),
                         dispformula=~trtGp+poly(TrtTime,2),
                         family=betabinomial(link="cloglog"), data=HawCon)
ablin2s.TMBclog <- update(ablin.TMBclog,
                          formula=cbind(Dead,Live)~0+
                          trtGp/TrtTime+poly(TrtTime,2)[,-1]+(TrtTime|gp))
ablin.TMB <- glmmTMB(cbind(Dead,Live)~0+trtGp/TrtTime+(TrtTime|gp),
                         dispformula=~trtGp+poly(TrtTime,2),
                         family=betabinomial(link="logit"), data=HawCon)
ablin2s.TMB <- update(ablin.TMB,
                         formula=cbind(Dead,Live)~0+trtGp/TrtTime+
                          poly(TrtTime,2)[,-1]+(TrtTime|gp))
```

The following shows AIC-based model comparisons:

```
                                          dAIC df
BB: Complementary log-log link             0.0 29
BB: Complementary log-log link, added curve  1.9 30
BB: Logit link, added curve                6.1 30
BB: Logit link                            17.1 29
```

In Figure 7.9A, unlike Figure 7.9B, the curves are close to straight lines. The comparison strongly favours a complementary log-log link, if there is no underlying smooth. The AIC based comparison points in the same direction. The complementary log-log link is strongly preferable to the logit if there is no smooth, with little difference between the smoothed and unsmoothed results in the former case.

We will now proceed to compare the fitted pattern of variation of the parameter ρ between the models with the two different link functions, when there is no underlying smooth.

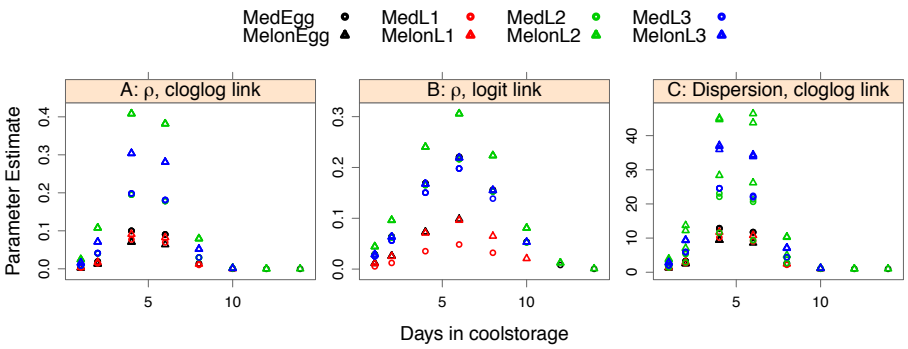

Figure 7.10: Panels A and B show intra-class correlation estimates for, respectively, a complementary log-log link and a logit link. Both models assume a beta binomial error.

Patterns of variation of intra-class correlation and of dispersion

Figures 7.10A (with a complementary log-log link) and B (with a logit link) show estimates of the intra-class correlation ρ for the model where an overall quadratic curve is added to a straight line response that is different for each different treatment group. Figure 7.10C shows the dispersion estimates that correspond to the estimates of ρ in Panel A.

Figure 7.10A plots the estimates of ρ, with the dispersion estimates in Panel B.

7.6.2 Diagnostic checks

Quantile residuals provide what can be effective checks. For any residual, the corresponding quantile residual is the proportion of residuals expected, under model assumptions, to be less than or equal to it. If the distributional assumptions are satisfied, the quantile residuals should have a distribution that differs only by statistical error from a uniform distribution. The function `DHARMa::simulateResiduals()` provides a convenient means to simulate enough sets of residuals to give a good sense, for each individual observation, of the distribution. These can then be used as a reference distribution for calculation of *quantile* residuals.

- For each observation, the quantile residual is the proportion of simulated residuals that are greater than the observed residual.
  - Thus, a value of 0 means that all simulated residuals are larger than the observed value, and a value of 0.5 means half of the simulated residuals are larger than the observed value.
- For a correctly specified model, the quantile residuals should be uniformly distributed on the unit interval.
  - For the data as a whole, this can be checked by plotting the quantile residuals against the corresponding quantiles.
  - A second check plots quantile residuals against quantiles of predicted values. Quantile regression is then used to fit curves at 25%, 50%, and 75% quantiles of the quantile residuals. If the model is correctly specified, these should all be, to within statistical error, horizontal lines.

Figure 7.11: Panel A shows quantile-quantile plots for the model and for three sets of simulated data. Panel B plots estimated quantiles against mortality, while Panel C plots estimated quantiles against total number, on a logarithmic scale.

We will use the function `DHARMa::simulateResiduals` to obtain repeated (by default, 250) simulations of residuals from the fitted model.
Figure 7.11 shows the diagnostic plots for the linear model with a complementary log-log link. These are then used to replace the residuals from the model with quantile residuals.

The following function `scaledSim()` will then be used to extract, also, several sets of quantile residuals that are derived from residuals that have been simulated from the model.

```
scaledSim <- function(simRef, nScale=1){
nObs <- simRef$nObs
nSim <- simRef$nSim
scaledRes <- matrix(0,nrow=nObs,ncol=nScale)
nSelect <- sample(nSim, size=nScale)
sampResponse <- simRef[['simulatedResponse']][ ,nSelect, drop=FALSE]
for(i in 1:nObs)scaledRes[i,] <- ecdf(simRef$simulatedResponse[i,] +
  runif(simRef$nSim, -0.5, 0.5))(sampResponse[i,] +
                                 runif(nScale, -0.5, 0.5))
scaledRes
}
```

The quantile-quantile plot (Q-Q) plot looks fine, The quantile residuals from the data appear, if anything, closer to uniformly distributed than any of the simulated sets of residuals. In Panels B and C, the quantile residuals from the data are well away from their respected reference horizontal lines, but are not obviously more so than the quantiles for the simulations.

7.6.3 Lethal time estimates and confidence intervals

The estimated time that is required to kill 99% of insects (lethal time 99, or LT99) is commonly used as a starting point for assessing what time might be effective in commercial practice. Thus, for the model that used a complementary log-log link, and setting:

$$y = \log(-\log(1-0.99)) = 1.53,$$

one solves for $x = \text{LT99}$ in an equation of the form $y = a + bx$. Thus:

$$\text{LT99} = x = \frac{y-a}{b}$$

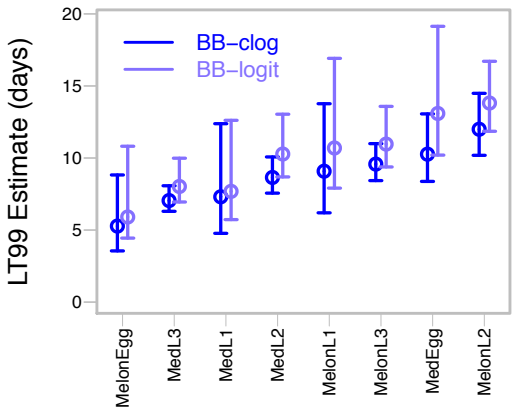

Figure 7.12: LT99 95 percent confidence intervals are compared between the model with a complementary log-log link, and the model with a logit link, in both cases with a beta-binomial error.

The determination of confidence intervals for such ratios is one of a much wider class of problems. The Fieller's formula approach (Morgan, 1992), implemented in the `qra` package, makes the common assumption that $(y-a, b)$ has been sampled from a distribution for which the bivariate normal is an adequate approximation. See `?qra::fieller`.

The sampling distribution of the calculated value x is typically, unless var[b] is small relative to b, long-tailed. As a consequence, the *Delta* method, which uses an approximate variance as the basis for inference, is in general unreliable. See `?qra::fieller`. The Fieller's formula approach cannot in general be adapted to give confidence intervals for differences of LT99s or other such ratio statistics, unless the denominators of the statistics that are compared happen to be the same. A usable implementation of the simulation approach, which seems needed for the calculation of confidence intervals for LT99 differences, requires adaptation of the function `lme4::bootMer()` to work with objects that have been fitted using `glmmTMB::glmmTMB`.

Figure 7.12 compares confidence intervals, calculated using Fieller's formula, from use of a complementary log-log link, with intervals from use of a logit link. In each case a two degrees of freedom polynomial was used to model the parameter that determines scale, and hence ρ.

Code to extract LT99 estimates and confidence intervals for the complementary log-log model is:

```
ablinLTClog <- qra::extractLT(p=0.99, obj=ablin.TMBclog, link="cloglog",
                              nEsts=8, slopeAdd=8, eps=0, df.t=NULL)[,-2]
rownames(ablinLTClog) <- shorten(rownames(ablinLTClog))
```

Other models that have been tried, but which give confidence intervals that are too wide to be shown satisfactorily on the same graph, are:

- A "Binomial" model with complementary log-log link ("Bin, clog").
- A "Binomial" model with logit link ("Bin, logit").
- A linear mixed model, with $log(1-log((p+0.002)/(1+0.004)))$ as the dependent variable, with complementary log-log link ("LMM, clog").

Code used is:

```
cloglog <- make.link('cloglog')$linkfun
abBINlin.TMBclog <- glmmTMB(cbind(Dead,Live)~0+trtGp/TrtTime+(scTime|gp),
                            family=binomial(link="cloglog"), data=HawCon)
abBINlin.TMB <- update(abBINlin.TMBclog, family=binomial(link="logit"))
```

Code to extract the LT99 estimates and confidence intervals for the first of the models is:

```
abBinLTClog <- qra::extractLT(p=0.99, obj=abBINlin.TMBclog, link="cloglog",
nEsts=8, slopeAdd=8, eps=0, scaling=c(1,1), df.t=NULL)[,-2]
rownames(abBinLTClog) <- shorten(rownames(abBinLTClog))
```

The following table gives the confidence intervals for the three models noted:

	Beta-binomial		Binomial				Transform, LMM	
	cloglog		cloglog		logit		cloglog transform	
MedEgg	**10.3**	**8.4 - 13.1**	**9.8**	**7.7 - 14.5**	**11.6**	**9 - 16.3**	**11.5**	**8.9 - 18.1**
MedL1	**7.3**	**4.8 - 12.4**	**6.4**	**3.4 - 11.7**	**6.9**	**4.9 - 10.4**	**7.3**	**3.9 - 20.4**
MedL2	8.7	7.6 - 10.1	8.1	7 - 9.6	9.8	8.3 - 12	8.6	7.6 - 10
MedL3	7	6.3 - 8.1	6.7	6.1 - 7.4	7.4	6.6 - 8.3	7.3	6.5 - 8.3
MelonEgg	**5.3**	**3.6 - 8.8**	**4.7**	**2.6 - 6.7**	**5.4**	**4.2 - 7.2**	**4.9**	**2.5 - 24.1**
MelonL1	**9.1**	**6.2 - 13.8**	**8.3**	**5 - 30.1**	**9.5**	**6.5 - 16.8**	**9.1**	**5.8 - 17.9**
MelonL2	12	10.2 - 14.5	10.4	8.2 - 15.6	11.3	9.1 - 14.9	11.5	9.9 - 14
MelonL3	9.6	8.4 - 11	9.2	7.8 - 11.4	10.5	8.8 - 13	9.6	8.5 - 11.1

Rows where one or more confidence interval upper bounds are high relative to those from the `glmmTMB` mode are highlighted in red. The first two columns, which show LT99s and confidence interval bounds from the `glmmTMB` model are likewise highlighted. With this particular dataset, it really is important to model the scale parameters. Models that do not account for dispersion effects that are a function of explanatory variables and/or factors will, with data of the type considered here, give confidence intervals for LT99s, or for other such statistics, that may be seriously in error.

7.7 Observation level random effects — the `moths` dataset

Consider again the `moths` data of Subsection 5.4.2. An alternative to use of a quasipoisson or negative binomial model to account for extra-Poisson variation is to allow for a random between transects error that is additive on the scale of the linear predictor. The model incorporates a term that allows for normally distributed random variation, additional to the Poisson variation at each observation. The model then incorporates what are termed "observation level random effects".

The attempt to fit a model that uses the default log link generates (lme4_1.1-7), if data for the habitat `Bank` is included, a warning that the model is nearly unidentifiable. Use of a square root link avoids this problem. Once again, `lowerside` will be used as reference level.