A Handbook of Statistical Analyses Using R — 2nd Edition

Brian S. Everitt and Torsten Hothorn

CHAPTER 14

Simultaneous Inference and Multiple Comparisons: Genetic Components of Alcoholism, Deer Browsing Intensities, and Cloud Seeding

14.1 Introduction

14.2 Simultaneous Inference and Multiple Comparisons

14.3 Analysis Using R

14.3.1 Genetic Components of Alcoholism

We start with a graphical display of the data. Three parallel boxplots shown in Figure 14.1 indicate increasing expression levels of alpha synuclein mRNA for longer *NACP*-REP1 alleles.

In order to model this relationship, we start fitting a simple one-way ANOVA model of the form $y_{ij} = \mu + \gamma_i + \varepsilon_{ij}$ to the data with independent normal errors $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$, $j \in \{\text{short}, \text{intermediate}, \text{long}\}$, and $i = 1, \ldots, n_j$. The parameters $\mu + \gamma_{\text{short}}$, $\mu + \gamma_{\text{intermediate}}$ and $\mu + \gamma_{\text{long}}$ can be interpreted as the mean expression levels in the corresponding groups. As already discussed in Chapter 5, this model description is overparameterised. A standard approach is to consider a suitable re-parameterization. The so-called "treatment contrast" vector $\theta = (\mu, \gamma_{\text{intermediate}} - \gamma_{\text{short}}, \gamma_{\text{long}} - \gamma_{\text{short}})$ (the default re-parameterization used as elemental parameters in R) is one possibility and is equivalent to imposing the restriction $\gamma_{\text{short}} = 0$.

In addition, we define all comparisons among our three groups by choosing $\mathbf{K}$ such that $\mathbf{K}\theta$ contains all three group differences (Tukey's all-pairwise comparisons):

$$\mathbf{K}_{\text{Tukey}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}$$

with parameters of interest

$$\vartheta_{\text{Tukey}} = \mathbf{K}_{\text{Tukey}}\theta = (\gamma_{\text{intermediate}} - \gamma_{\text{short}}, \gamma_{\text{long}} - \gamma_{\text{short}}, \gamma_{\text{long}} - \gamma_{\text{intermediate}}).$$

The function `glht` (for generalised linear hypothesis) from package **multcomp** (Hothorn et al., 2009, 2008) takes the fitted *aov* object and a description

```
R> n <- table(alpha$alength)
R> levels(alpha$alength) <- abbreviate(levels(alpha$alength), 4)
R> plot(elevel ~ alength, data = alpha, varwidth = TRUE,
+       ylab = "Expression Level",
+       xlab = "NACP-REP1 Allele Length")
R> axis(3, at = 1:3, labels = paste("n = ", n))
```

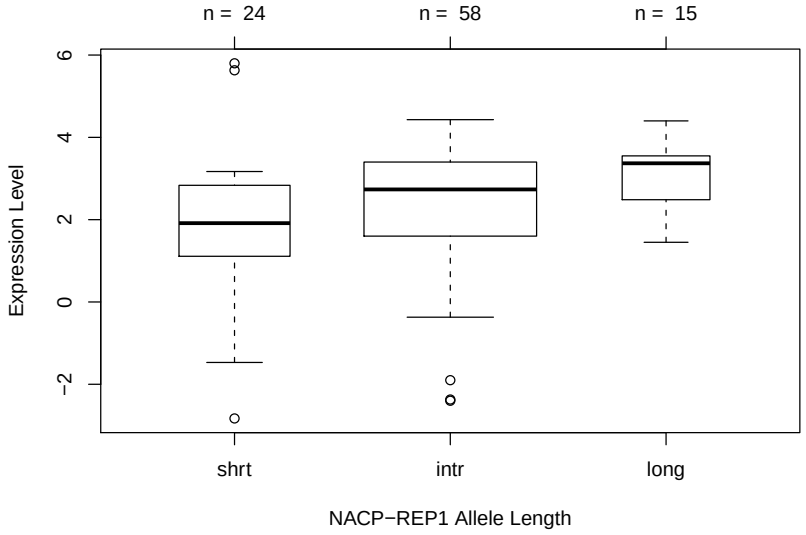

Figure 14.1 Distribution of levels of expressed alpha synuclein mRNA in three groups defined by the *NACP*-REP1 allele lengths.

of the matrix **K**. Here, we use the `mcp` function to set up the matrix of all pairwise differences for the model parameters associated with factor `alength`:

```
R> library("multcomp")
R> amod <- aov(elevel ~ alength, data = alpha)
R> amod_glht <- glht(amod, linfct = mcp(alength = "Tukey"))
```

The matrix **K** reads

```
R> amod_glht$linfct
```

```
             (Intercept) alengthintr alengthlong
intr - shrt            0           1           0
long - shrt            0           0           1
long - intr            0          -1           1
attr(,"type")
[1] "Tukey"
```

The `amod_glht` object now contains information about the estimated linear

function $\hat{\vartheta}$ and their covariance matrix which can be inspected via the `coef` and `vcov` methods:

```
R> coef(amod_glht)
```

```
intr - shrt long - shrt long - intr
      0.434       1.189       0.755
```

```
R> vcov(amod_glht)
```

```
            intr - shrt long - shrt long - intr
intr - shrt      0.1472       0.104     -0.0431
long - shrt      0.1041       0.271      0.1666
long - intr     -0.0431       0.167      0.2096
```

The `summary` and `confint` methods can be used to compute a summary statistic including adjusted p-values and simultaneous confidence intervals, respectively:

```
R> confint(amod_glht)
```

```
         Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = elevel ~ alength, data = alpha)

Estimated Quantile = 2.37
95% family-wise confidence level


Linear Hypotheses:
                 Estimate lwr     upr
intr - shrt == 0  0.4342  -0.4757  1.3440
long - shrt == 0  1.1887  -0.0452  2.4227
long - intr == 0  0.7546  -0.3313  1.8405
```

```
R> summary(amod_glht)
```

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = elevel ~ alength, data = alpha)

Linear Hypotheses:
                 Estimate Std. Error t value Pr(>|t|)
intr - shrt == 0    0.434      0.384    1.13    0.492
long - shrt == 0    1.189      0.520    2.28    0.061
long - intr == 0    0.755      0.458    1.65    0.227
(Adjusted p values reported -- single-step method)
```

Because of the variance heterogeneity that can be observed in Figure 14.1, one might be concerned with the validity of the above results stating that there is no difference between any combination of the three allele lengths. A sandwich estimator might be more appropriate in this situation, and the `vcov` argument can be used to specify a function to compute some alternative covariance estimator as follows:

```
R> amod_glht_sw <- glht(amod, linfct = mcp(alength = "Tukey"),
+                       vcov = sandwich)
R> summary(amod_glht_sw)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = elevel ~ alength, data = alpha)

Linear Hypotheses:
                  Estimate Std. Error t value Pr(>|t|)
intr - shrt == 0    0.434      0.424    1.02    0.559
long - shrt == 0    1.189      0.443    2.68    0.023
long - intr == 0    0.755      0.318    2.37    0.050
(Adjusted p values reported -- single-step method)
```

We use the `sandwich` function from package **sandwich** (Zeileis, 2004, 2006) which provides us with a heteroscedasticity-consistent estimator of the covariance matrix. This result is more in line with previously published findings for this study obtained from non-parametric test procedures such as the Kruskal-Wallis test. A comparison of the simultaneous confidence intervals calculated based on the ordinary and sandwich estimator is given in Figure 14.2.

It should be noted that this data set is heavily unbalanced; see Figure 14.1, and therefore the results obtained from function `TukeyHSD` might be less accurate.

14.3.2 Deer Browsing

Since we have to take the spatial structure of the deer browsing data into account, we cannot simply use a logistic regression model as introduced in Chapter 7. One possibility is to apply a mixed logistic regression model (using package **lme4**, Bates and Sarkar, 2008) with random intercept accounting for the spatial variation of the trees. These models have already been discussed in Chapter 13. For each plot nested within a set of five plots oriented on a 100m transect (the location of the transect is determined by a predefined equally spaced lattice of the area under test), a random intercept is included in the model. Essentially, trees that are close to each other are handled like repeated measurements in a longitudinal analysis. We are interested in probability estimates and confidence intervals for each tree species. Each of the six fixed

```
R> par(mai = par("mai") * c(1, 2.1, 1, 0.5))
R> layout(matrix(1:2, ncol = 2))
R> ci1 <- confint(glht(amod, linfct = mcp(alength = "Tukey")))
R> ci2 <- confint(glht(amod, linfct = mcp(alength = "Tukey"),
+                 vcov = sandwich))
R> ox <- expression(paste("Tukey (ordinary ", bold(S)[n], ")"))
R> sx <- expression(paste("Tukey (sandwich ", bold(S)[n], ")"))
R> plot(ci1, xlim = c(-0.6, 2.6), main = ox,
+      xlab = "Difference", ylim = c(0.5, 3.5))
R> plot(ci2, xlim = c(-0.6, 2.6), main = sx,
+      xlab = "Difference", ylim = c(0.5, 3.5))
```

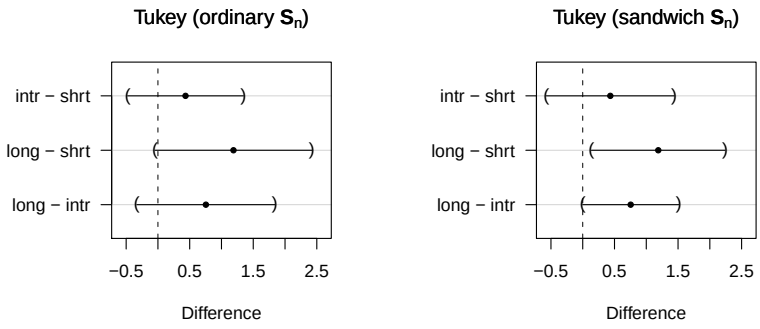

Figure 14.2 Simultaneous confidence intervals for the `alpha` data based on the ordinary covariance matrix (left) and a sandwich estimator (right).

parameters of the model corresponds to one species (in absence of a global intercept term); therefore, $\mathbf{K} = \text{diag}(6)$ is the linear function we are interested in:

```
R> mmod <- lmer(damage ~ species - 1 + (1 | lattice / plot),
+               data = trees513, family = binomial())
R> K <- diag(length(fixef(mmod)))
R> K
```

```
     [,1] [,2] [,3] [,4] [,5]
[1,]    1    0    0    0    0
[2,]    0    1    0    0    0
[3,]    0    0    1    0    0
[4,]    0    0    0    1    0
[5,]    0    0    0    0    1
```

In order to help interpretation, the names of the tree species and the corresponding sample sizes (computed via `table`) are added to $\mathbf{K}$ as row names; this information will carry through all subsequent steps of our analysis:

```
R> colnames(K) <- rownames(K) <-
+      paste(gsub("species", "", names(fixef(mmod))),
+            " (", table(trees513$species), ")", sep = "")
R> K
```

```
                spruce (119) pine (823) beech (266) oak (1258)
spruce (119)               1          0           0          0
pine (823)                 0          1           0          0
beech (266)                0          0           1          0
oak (1258)                 0          0           0          1
hardwood (191)             0          0           0          0
                hardwood (191)
spruce (119)                 0
pine (823)                   0
beech (266)                  0
oak (1258)                   0
hardwood (191)               1
```

Based on $\mathbf{K}$, we first compute simultaneous confidence intervals for $\mathbf{K}\theta$ and transform these into probabilities. Note that $\left(1 + \exp(-\hat{\vartheta})\right)^{-1}$ (cf. Equation **??**) is the vector of estimated probabilities; simultaneous confidence intervals can be transformed to the probability scale in the same way:

```
R> ci <- confint(glht(mmod, linfct = K))
R> ci$confint <- 1 - binomial()$linkinv(ci$confint)
R> ci$confint[,2:3] <- ci$confint[,3:2]
```

The result is shown in Figure 14.3. Browsing is less frequent in hardwood but especially small oak trees are severely at risk. Consequently, the local authorities increased the number of roe deers to be harvested in the following years. The large confidence interval for ash, maple, elm and lime trees is caused by the small sample size.

14.3.3 Cloud Seeding

In Chapter 6 we studied the dependency of rainfall on S-Ne values by means of linear models. Because the number of observations is small, an additional assessment of the variability of the fitted regression lines is interesting. Here, we are interested in a confidence band around some estimated regression line, i.e., a confidence region which covers the true but unknown regression line with probability greater or equal $1 - \alpha$. It is straightforward to compute *pointwise* confidence intervals but we have to make sure that the type I error is controlled for all x values simultaneously. Consider the simple linear regression model

$$\text{rainfall}_i = \beta_0 + \beta_1 \text{sne}_i + \varepsilon_i$$

where we are interested in a confidence band for the predicted rainfall, i.e., the values $\hat{\beta}_0 + \hat{\beta}_1 \text{sne}_i$ for some observations sne_i. (Note that the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables.)

```
R> plot(ci, xlab = "Probability of Damage Caused by Browsing",
+       xlim = c(0, 0.5), main = "", ylim = c(0.5, 5.5))
```

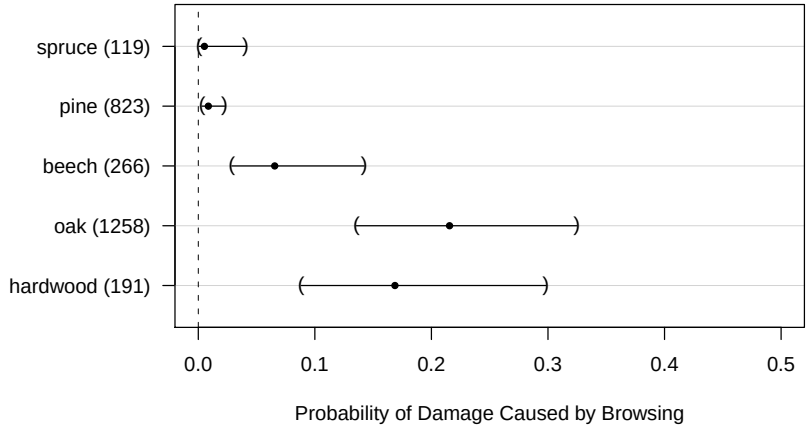

Figure 14.3 Probability of damage caused by roe deer browsing for six tree species. Sample sizes are given in brackets.

We can formulate the problem as a linear combination of the regression coefficients by multiplying a matrix $\mathbf{K}$ to a grid of S-Ne values (ranging from 1.5 to 4.5, say) from the left to the elemental parameters $\theta = (\beta_0, \beta_1)$:

$$\mathbf{K}\theta = \begin{pmatrix} 1 & 1.50 \\ 1 & 1.75 \\ \vdots & \vdots \\ 1 & 4.25 \\ 1 & 4.50 \end{pmatrix} \theta = (\beta_0 + \beta_1 1.50, \beta_0 + \beta_1 1.75, \ldots, \beta_0 + \beta_1 4.50) = \vartheta.$$

Simultaneous confidence intervals for all the parameters of interest ϑ form a confidence band for the estimated regression line. We implement this idea for the `clouds` data writing a small reusable function as follows:

```
R> confband <- function(subset, main) {
+     mod <- lm(rainfall ~ sne, data = clouds, subset = subset)
+     sne_grid <- seq(from = 1.5, to = 4.5, by = 0.25)
+     K <- cbind(1, sne_grid)
+     sne_ci <- confint(glht(mod, linfct = K))
+     plot(rainfall ~ sne, data = clouds, subset = subset,
+          xlab = "S-Ne criterion", main = main,
+          xlim = range(clouds$sne),
```

```
R> layout(matrix(1:2, ncol = 2))
R> confband(clouds$seeding == "no", main = "No seeding")
R> confband(clouds$seeding == "yes", main = "Seeding")
```

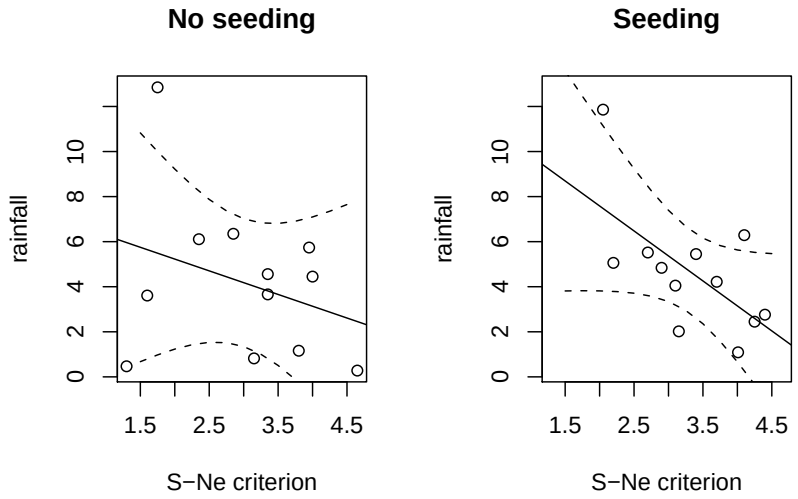

Figure 14.4 Regression relationship between S-Ne criterion and rainfall with and without seeding. The confidence bands cover the area within the dashed curves.

```
+           ylim = range(clouds$rainfall))
+      abline(mod)
+      lines(sne_grid, sne_ci$confint[,2], lty = 2)
+      lines(sne_grid, sne_ci$confint[,3], lty = 2)
+  }
```

The function `confband` basically fits a linear model using `lm` to a subset of the data, sets up the matrix $\mathbf{K}$ as shown above and nicely plots both the regression line and the confidence band. Now, this function can be reused to produce plots similar to Figure **??** separately for days with and without cloud seeding in Figure 14.4. For the days without seeding, there is more uncertainty about the true regression line compared to the days with cloud seeding. Clearly, this is caused by the larger variability of the observations in the left part of the figure.

14.4 Summary

Multiple comparisons in linear models have been in use for a long time. The **multcomp** package extends much of the theory to a broad class of parametric

and semi-parametric statistical models, which allows for a unified treatment of multiple comparisons and other simultaneous inference procedures in generalised linear models, mixed models, models for censored data, robust models, etc. Honest decisions based on simultaneous inference procedures maintaining a pre-specified familywise error rate (at least asymptotically) can be derived from almost all classical and modern statistical models. The technical details and more examples can be found in Hothorn et al. (2008) and the package vignettes of package **multcomp** (Hothorn et al., 2009).

Exercises

Ex. 14.1 Compare the results of `glht` and `TukeyHSD` on the `alpha` data.

Ex. 14.2 Consider the linear model fitted to the clouds data as summarised in Figure **??**. Set up a matrix **K** corresponding to the global null hypothesis that all interaction terms present in the model are zero. Test both the global hypothesis and all hypotheses corresponding to each of the interaction terms. Which interaction remains significant after adjustment for multiple testing?

Ex. 14.3 For the logistic regression model presented in Figure **??** perform a multiplicity adjusted test on all regression coefficients (except for the intercept) being zero. Do the conclusions drawn in Chapter 7 remain valid?

Bibliography

Bates, D. and Sarkar, D. (2008), ***lme4****: Linear Mixed-Effects Models Using S4 Classes*, URL `http://CRAN.R-project.org/package=lme4`, R package version 0.999375-28.

Hothorn, T., Bretz, F., and Westfall, P. (2008), "Simultaneous inference in general parametric models," *Biometrical Journal*, 50, 346–363.

Hothorn, T., Bretz, F., and Westfall, P. (2009), ***multcomp****: Simultaneous Inference for General Linear Hypotheses*, URL `http://CRAN.R-project.org/package=multcomp`, R package version 1.0-7.

Zeileis, A. (2004), "Econometric computing with HC and HAC covariance matrix estimators," *Journal of Statistical Software*, 11, 1–17, URL `http://www.jstatsoft.org/v11/i10/`.

Zeileis, A. (2006), "Object-oriented computation of sandwich estimators," *Journal of Statistical Software*, 16, 1–16, URL `http://www.jstatsoft.org/v16/i09/`.