Lecture 2
Application of Linear Models

Chao Song

College of Ecology

Lanzhou University

July 16, 2025

Statistical issues in applying linear models

Using linear model for data analysis involves much more than just fitting the model using least squares. These include:

- Model interpretation;
- Complex hypothesis testing;
- Model diagnostics;
- Model building

The meaning of model parameters

In linear models, there are many ways to parameterize categorical predictors. This is commonly referred to as contrast coding in statistical software. Common contrasting codings include reference level coding, sum to zero coding, Helmert coding.

Different contrasting coding are statistically equivalent in terms of model fitting, but they result in different meaning of parameter estimates. Choose a particular coding scheme to suit your need of analysis.

Contrast coding

Example: For a simple one way ANOVA model with k groups, we have the model

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}.$$

- With reference level coding, μ is the mean of a chosen reference level and α_i is the difference between other levels and the reference.
- With sum to zero coding, μ is the overall mean and α_i represents the deviation of each group from the grand mean.

Hypothesis testing in linear models

Methods of least squares and the assumption of independent and identically distributed normal errors lead to F test for linear hypotheses in linear models.

- F-test as implemented in function `lht()` in package `car` is a general way for any linear hypothesis testing.
- Function exists to generate ANOVA tables for testing existence of effects as this is the most common type of hypothesis test in using linear model. Be aware that sum of squares can be calculated in different ways. Choose the appropriate one for your hypothesis.

Type of sum of squares

Type I SS for a given factor represents the reduction in error sum of squares by going from a model without that factor to a model with that factor. With Type I SS, decomposition of total sum of squares is guaranteed to hold, but the effects of factor often depends on the order it enters the model.

Type II SS for a given factor is computed as the reduction in SS by adding the term of the factor to a model that is the largest hierarchical model that does not contain that focal factor. Type II SS is order independent.

Type III SS computes the sum of square for a factor by adding that factor to the model with all other potential factors. ANOVA table constructed from Type III SS lead to testing whether marginal means are equal.

Type of sum of squares

Some comments on the type of sum of squares:

- In terms of the marginal means, the hypotheses tested by type I and II SS are difficult to interpret. Type III SS tests whether group means are equal. Thus, in most cases, we should chose type III SS in constructing ANOVA table.
- For balanced design, all three types of SS gives the same results. For unbalanced designed, they do not.
- When using type III SS, factors should always be coded by the sum to zero coding.
- The `anova()` function in base R uses type I SS. Function `Anova()` in package `car` implement type II and III ss.

Multiple comparisons

Once a particular hypothesis is rejected, we often want to further test which group differs from which. This raises the multiple comparisons problem.

- When performing a single statistical hypothesis test, we try to avoid incorrectly rejecting the null hypothesis for that test by setting this probability of such an error to be low;
- When conducting multiple hypothesis tests, the probability of making at least one type I error increases the more tests we perform.
- We thus need to adjust how we test each individual hypothesis tests so that we control the family wise type I error rate, i.e., the probability of making at least one type I error among all inferences in a family.

Multiple comparisons

There is no universally agreed upon recommendation on which method to choose. Some general suggestions:

- Use Tukey HSD method for all pairwise comparisons;
- Use Dunnett’s method for all pairwise comparisons with a single reference mean;
- Use Bonderroni method for other simultaneous multiple comparisons
- Use Fisher‘s LSD for other planned comparisons

Model diagnostics

Statistical inference in linear models arises from the assumptions of linearity and independent and identically distributed normal errors. Thus, we need to check whether these assumptions are met. This is typically done by examining residuals of the model. Check for:

- unequal spread of residual
- non-independence of residuals
- normality of residuals
- outliers or highly influential point

Although there are formal statistical test to test these assumptions, they typically have low power. Thus, my recommendation is to just visually examine the residual plots.

Handling violation of assumptions

If we find evidence for violation of model assumptions, we can typically

- Revising model structures such as adding additional predictors, changing forms of the model.
- Transforming the response variable

A common family of transformation is the **Box-Cox** transformation:

$$f(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, \ \lambda \neq 0; \\ \log(y), \ \lambda = 0 \end{cases}$$

Here, λ is estimated by maximum likelihood estimation. Roughly speaking, we find the value of λ such that the transformed response variable best fit to the model assumption.