Tutorial 3

STAT3015/4030/7030 Generalised Linear Modelling

The Australian National University

Week 3, 2017

Overview

1. Summary
   - One-way ANOVA
   - Two-way ANOVA
2. Question 2
3. Question 3

One-way ANOVA Model

We denote sampled data values as Y_{ij}, where $i = 1, \ldots, k$ indicates the factor level and $j = 1, \ldots, n_i$ indicates a specific value within the i^{th} factor level. We might write:

$$Y_{ij} = \mu + \tau_i + \varepsilon_i,$$

with some constraints to avoid overparameterisation. Here τ_i is the i^{th} level effect or treatment effect.

- Treatment contrasts. $\tau_1 = 0$
- Sum contrasts. $\sum_{i=1}^{k} n_i \tau_i = 0$

Contrast of μ_i's

We can find a 100(1-α)% confidence interval for any linear combination of the μ_i's, say $h_1\mu_1 + \cdots + h_k\mu_k$, for any vector of constants $h = (h_1, \ldots, h_k)$. Such a linear combination is often called a contrast.

Since normally "within factor" averages are formed from disjoint (and therefore independent) subsets of the observed responses, we have $\bar{Y}_i$'s are independent. Then we have

$$Var(\sum_{i=1}^{k} h_i \bar{Y}_i) = \sum_{i=1}^{k} h_i^2 Var(\bar{Y}_i) = \sigma^2 \sum_{i=1}^{k} \frac{h_i^2}{n_i}.$$

Contrast of μ_i's

Thus, the desired confidence interval would be

$$(\sum_{i=1}^{k} h_i \bar{Y}_i) \pm t_{n-k}(1-\frac{\alpha}{2}) s \sqrt{\sum_{i=1}^{k} \frac{h_i^2}{n_i}}.$$

We can also test hypotheses of the form:

$$H_0 : \sum_{i=1}^{k} h_i \mu_i = c_0 \quad versus \quad H_0 : \sum_{i=1}^{k} h_i \mu_i \neq c_0.$$

Using the test statistic:

$$T = \frac{\sum_{i=1}^{k} h_i \bar{Y}_i - c_0}{s\sqrt{\sum_{i=1}^{k} \frac{h_i^2}{n_i}}}.$$

Two-way ANOVA model

Two-way ANOVA model is appropriate for datasets that contain a continuous numerical response variable and two categorical predictors.

Y_{ijk} means the k^{th} measurement observed at the i^{th} ($k = 1, \cdots, n$) level of the first factor ($i = 1, \cdots, I$) and the j^{th} factor of the second factor ($j = 1, \cdots, J$). With a balanced design, we have the additive model

$$Y_{ijk} = \mu_i + \nu_j + \epsilon_{ijk} = \mu + \tau_i + \alpha_j + \epsilon_{ijk},$$

where $\mu_i + \nu_j$ is the expected response within the $(i, j)^{th}$ level combination of the two factors, μ_i representing the effect on the expected response of the i^{th} level of the first factor and ν_j the effect of the j^{th} level of the second factor.

Two-way ANOVA model

The previous model assumes that the effects of the two factors are additive: the effect of the either factor is not changed depending on the level of the other factor at which the observations are being made.

No interaction between two factors!

Two sets of commonly used constraints:

- the "baseline" or "control group structure": $\tau_1 = \alpha_1 = 0$; or,
- the "grand mean" constraints: $\sum_{i=1}^{I} \tau_i = \sum_{j=1}^{J} \alpha_j = 0$.

We can still use indicator variables and multiple regression techniques to write the model using matrix notation as:

$$\boldsymbol{Y} = \boldsymbol{1_n}\boldsymbol{\beta_0} + \boldsymbol{Z}\boldsymbol{\beta_{(1)}} + \boldsymbol{W}\boldsymbol{\beta_{(2)}} + \boldsymbol{\varepsilon}$$

Two-way ANOVA model

We still use the sequential F-statistic to do the hypothesis test. For example,

$$H_0 : \beta_{(2)} = 0,$$

$$F = \frac{SSR(\beta_{(2)}|\beta_{(1)}, \beta_0)/(J-1)}{MSE_{full}}$$

which has an F-distribution with $J-1$ numerator and $nIJ - (I + J - 1)$ denominator degrees of freedom.

(in analogy to testing of a subset of βs in multiple linear regression)

- Input data and create random variables. (Anyone needs some help?)
- Fit two-way anova models and refit the model using indicator variables.
- Multiply indicators to get the two-factor interaction. Include the interaction term last in the linear model.

- Part (a) is similar to Q2 (testing of interaction term)
- For (b), use summary output to find estimated coefficients and test their significance. (partial T-test)

- Part (a) is similar to Q2 (testing of interaction term)
- For (b), use summary output to find estimated coefficients and test their significance. (partial T-test)
- For (b) we can also use `drop1(model_name, test="F")` to test the significance of one factor. (after accounting for the association with the other factor)
- For (b) we can do formal test for equality of the regression coefficients similar to what we have done in Tutorial 2. (create constant vector `h=c(0,-1,1)` and do matrix multiplication; find CI and p-value)