Tutorial 2

STAT3015/4030/7030 Generalised Linear Modelling

The Australian National University

Week 2, 2017

Overview

Linear Model

The general form of linear models is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

The above equation can be written in a matrix/vector representation as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T, \boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T, \boldsymbol{\beta} = (\beta_0, \ldots, \beta_n)^T$ and $\mathbf{X}$ is the design matrix.

Least Squares Estimation

We define the best estimate of $\boldsymbol{\beta}$ as the one which minimizes the sum of the squared errors:

$$\sum \varepsilon_i^2 = \boldsymbol{\varepsilon}^{\boldsymbol{T}} \boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\beta)^T (\boldsymbol{y} - \boldsymbol{X}\beta)$$

Differentiating with respect to $\boldsymbol{\beta}$ and setting to zero, we find that $\hat{\boldsymbol{\beta}}$ satisfies:

$$\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X} \hat{\beta} = \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y}$$

Therefore,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{X})^{-1} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{y}$$

Inference

By using the least squares estimation we have assumed that the errors are independent and identically distributed (i.i.d.) with mean 0 and variance $\boldsymbol{\sigma^2}$, so we have

$$\varepsilon \sim N(\mathbf{0}, \boldsymbol{\sigma}^2 \boldsymbol{I})$$

Since $\boldsymbol{y} = \boldsymbol{X}\beta + \varepsilon$, we have

$$\boldsymbol{y} \sim N(\boldsymbol{X}\beta, \boldsymbol{\sigma}^2 \boldsymbol{I})$$

Using the fact that linear combinations of normally distributed values are also normal, we find that:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \sim N(\beta, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{\sigma}^2)$$

Can you calculate a 100(1-α)% CI for $\hat{\boldsymbol{\beta}}$?

Hypothesis Tests

If $\boldsymbol{X}$ is a $n \times p$ matrix, we can conduct an overall test of the model under

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0$$

by referring to $F_{p-1,n-p}$.
We can also test the significance of each predictor under

$$H_0 : \beta_i = 0$$

by using a t-statistic

$$t_i = \hat{\beta}_i / se(\hat{\beta}_i)$$

How to test $H_0 : \beta_i = constant \times \beta_j$? (Question 1 (c))

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. $\sum_{i=1}^{k} n_i \tau_i = 0$

The two parameterisations have different formats of estimators of μ_i and τ_i (Page 3-5 of Lecture Brick).

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:

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

Question 1 (c) and Question 2 (b) & (c)

- The main difference of `aov()` from `lm()` is in the way `print`, `summary` and so on handle the fit.
- `aov()` is designed for balanced designs. **Is this question a balanced experiment?**
- For (c), we need to firstly find a vector of constant h. Then we can use a similar to the "corn yield" example (Page 7 of Brick), OR, consider

$$\boldsymbol{h^T}\hat{\beta} \sim \boldsymbol{N}(\boldsymbol{h^T}\beta, \boldsymbol{h^T}(\boldsymbol{X^T X})^{-1}\boldsymbol{h}\boldsymbol{\sigma}^2)$$

- s, which is the estimator of σ, can be find by `summary(model)$sigma`. This method only works for models created by `lm()`.

Some hints

- There is no dataset provided for this question. We need to manually input the data using `c()`.
- To find the level means, we can use `tapply(values, factor, mean)`.
- For `aov()` model, we calculate

$$s^2 = MSE = \frac{SSE}{n-k},$$

  where $SSE = \sum(Observed - Fitted)^2$
- For the second part of (c), we can firstly simplify the Null hypothesis before calculating test statistic.