STA610 Lab01 ANOVA and REANOVA

Hun Kang

2024-08-30

- Write down your answers in any blank sheet and submit your work in paper during the lab.
- Your work will not be graded. As long as you submit, you will get a full credit.
- For those who missed the lab today, you can submit it via email to me for half credit.

1. ANOVA as a linear regression

Consider an one-way ANOVA model where an ith unit of a jth group is modeled as

$$\begin{aligned} y_{ij} &= \theta_j + \epsilon_{ij}, \quad i \in [n],\ j \in [m] \\ \theta_j &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

Another way to look at this model is as follows. First, we reparameterize the group mean parameters θ_j as

$$\theta_j = \begin{cases} \alpha & j = 1 \\ \alpha + \beta_{j-1} & j \geq 2 \end{cases}$$

Then you can check that the above model can be written as

$$\boldsymbol{y} = \alpha 1_{mn} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where 1_{mn} is a vector of 1 of length mn, $\boldsymbol{y} = [y_{11}, \cdots, y_{nm}]^T$, similarly for $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_{m-1}]^T$.

Q1-1: Write down the form of a design matrix $\boldsymbol{X}$.

The hypothesis of testing no difference in group means can be written as

$$H_0 : \theta_1 = \cdots = \theta_p \quad \text{vs} \quad H_1 : \theta_i \neq \theta_j \ \ ^{\exists}(i, j)$$

Q1-2: Re-express the above H_0 using β.

This tells us that ANOVA can be seen as a linear regression, and its hypothesis testing is equivalent to testing a submodel of linear regression. We do not proceed from here, but the general idea is as follows. Note that we wrote ANOVA decomposition as $SST = SSA + SSW$ where

$$\begin{aligned} SST &= \sum_{j=1}^{m}\sum_{i=1}^{n} (y_{ij} - \bar{y})^2 \\ SSA &= \sum_{j=1}^{m}\sum_{i=1}^{n} (\bar{y}_j - \bar{y})^2 \\ SSW &= \sum_{j=1}^{m}\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \end{aligned}$$

For a matrix X, we write $c(X)$ the vector space its columns expand, i.e., column space. Also, we write P_X an orthogonal projection matrix onto $c(X)$. For convenience, let $Z = cbind(1_{mn}, X)$. The dimension of $c(Z)$ is m.

1. One can see that $SST = \|(I - P_1)y\|^2 = y^T(I - P_1)y$, $SSA = y^T(P_Z - P_1)y$ and $SSW = y^T(I - P_Z)y$.
2. $(I - P_1)y$ is a vector y projected onto the column space orthogonal to $c(1_{mn})$, whose dimension is $mn - 1$.
3. $(I - P_Z)y$ is a vector y projected onto the column space orthogonal to $c(Z)$, whose dimension is $mn - m = m(n - 1)$.
4. $(P_Z - P_1)y$ is a vector y projected onto the subspace of $c(Z)$ that are orthogonal to $c(1_{mn})$, whose dimension is $m - 1$

Using these, the F-statistics of testing H_0 is

$$F(y) = \frac{SSA/(m-1)}{SSW/m(n-1)}$$

Note that $SSA = y^T(P_Z - P_1)y = y^T(I - P_1)y - y^T(I - P_Z)y$. A different interpretation of $y^T(I - P_Z)y$ is to see it as a residual sum of squares of a model with a design matrix Z. In this aspect, large SSA means that by including X, we see a large decrease in the residuals, i.e., the full model with Z fits the data better than the intercept only model.

```r
library(tidyverse)
URL <- "https://campus.murraystate.edu/academic/faculty/cmecklin/STA565/wheat.txt"
wheat <- read.table(URL,header=TRUE)
```

```r
str(wheat)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ variety : chr  "A" "A" "A" "A" ...
##  $ location: int  1 2 3 4 5 6 1 2 3 4 ...
##  $ yield   : num  35.3 31 32.7 36.8 37.2 33.1 33.7 32.2 31.4 32.7 ...
```

```r
lm1 = lm(yield ~ 1, data = wheat)
lm2 = lm(yield ~ 1 + as.factor(location), data = wheat)
anova(lm1, lm2)
```

```
## Analysis of Variance Table
##
## Model 1: yield ~ 1
## Model 2: yield ~ 1 + as.factor(location)
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     29 151.34
## 2     24 103.60  5    47.742 2.212 0.08631 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> **Q1-3: Conduct the f-test of treatment effect as learned in the class and check that the test statistics is the same.**

2. REANOVA and covariance structure

Consider a REANOVA model

$$\begin{aligned} y_{ij} &= \mu + a_j + \epsilon_{ij}, \quad i \in [n_j],\ j \in [m]\ (n_j = n\ {}^\forall j) \\ a_j &\overset{iid}{\sim} N(0, \tau^2) \\ \epsilon_{ij} &\overset{iid}{\sim} N(0, \sigma^2), \quad E[a_j \epsilon_{ij}] = 0 \end{aligned}$$

where μ, σ^2 and τ^2 are some unknown fixed parameter.

In the class, we saw that $E(y_{ij}) = \mu$, $V(y_{ij}) = \sigma^2 + \tau^2$ and $Cov(y_{1j}, y_{2j}) = \tau^2$. Also, since y_{ij} is a sum of Gausssian random variables, it itself also follows normal distribution. From this, we can write a joint distribution of $\boldsymbol{y}_j = [y_{1j}, \cdots, y_{nj}]^T$ for a group j as

$$\boldsymbol{y}_j \sim N(\mu 1_n, \Sigma_j)$$

Q2-1: Write down the covariance matrix Σ_j as follows:

$$(\Sigma_j)_{kl} = \begin{cases} ? & (k = l) \\ ? & (k \neq l) \end{cases}$$

Combining all $\boldsymbol{y}_j$, we can rewrite the above REANOVA model in a matrix-vector form as

$$\boldsymbol{y} \sim N(\mu 1_{mn}, \Sigma)$$

Q2-2: What is $Cov(y_{i1}, y_{i2})$? Using this, how can we write Σ?

In contrast, if we change the model as

$$\begin{aligned} y_{ij} &= \mu + \alpha_j + \epsilon_{ij}, \quad i \in [n_j],\ j \in [m]\ (n_j = n\ {}^\forall j) \\ \epsilon_{ij} &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

where α_j is also some unknown fixed parameter, then we have

$$\boldsymbol{y}_j \sim N\left((\mu + \alpha_j) 1_n, \sigma^2 I_n\right)$$

Q2-3: Write down the joint model for $\boldsymbol{y}$

The key takeaway is that, by adding another source of randomness a_j for group-wise variation, marginally we are modelling the covariance structure of $\boldsymbol{y}$ as Σ.