Supplementary Note 2

Unbalanced ANOVA and Regression Approach

One-factor ANOVA model is

$$Y_{ij} = \mu_i + \varepsilon_{ij}, j = 1, \ldots, n_i, i = 1, \ldots, k,$$

where $\{\varepsilon_{ij}\}$ are iid $N(0, \sigma^2)$. In order to write this as a factor effects model, we write

$$\mu_i = \mu_. + (\mu_i - \mu_.) := \mu_. + \alpha_i,$$

where $\mu_. = w_1\mu_1 + \cdots + w_k\mu_k$ is a weighted average of $\{\mu_i\}$. The choice of the weights $\{w_i\}$ is ours.

The usual choice of weights in the one-factor case is $w_i = n_i/n_T$, where n_T is the total number of observations. However, we may also choose equal weights $w_i = 1/k$. For any choice of weight $\{w_i\}$, the factor effects $\{\alpha_i\}$ satisfy the constraint

$$\sum w_i \alpha_i = 0.$$

Thus we have

$$\begin{aligned} \sum (n_i/n_T)\alpha_i &= 0 \text{ if } w_i = n_i/n_T, \text{ and} \\ \sum (1/k)\alpha_i &= 0 \text{ if } w_i = 1/k. \end{aligned}$$

Since $\{\bar{Y}_{i.}\}$ are the least squares estimate of $\{\mu_i\}$, the least squares estimates of $\mu_.$ and α_i are

$$\hat{\mu}_. = \sum w_i \bar{Y}_{i.}, \; \hat{\alpha}_i = \bar{Y}_{i.} - \hat{\mu}_..$$

It also follows that

$$\sum w_i \hat{\alpha}_i = 0.$$

Thus when $w_i = n_i/n_T$, we have

$$\hat{\mu}_. = \sum (n_i/n_T)\bar{Y}_{i.} = \bar{Y}_{..}, \; \hat{\alpha}_i = \bar{Y}_{i.} - \hat{\mu}_. = \bar{Y}_{i.} - \bar{Y}_{..}.$$

When $w_i = 1/k$, we have

$$\hat{\mu}_. = \sum (1/n)\bar{Y}_{i.} = \bar{Y}_{..}, \; \hat{\alpha}_i = \bar{Y}_{i.} - \hat{\mu}_..$$

For the balances case, the two weighing schemes, i.e., $\{w_i = n_i/n_T\}$ and $\{w_i = 1/k\}$, are the same and they lead to the same estimates of $\mu_.$ and α_i. are equal and thus the estimates of $\mu_.$ and α_i are identical. However, the estimates are not the same for the unbalanced case.

A Simple Regression Approach.

For the factor effects model we may use a simple regression approach to rewrite the ANOVA model as a regression model. Define $k-1$ indicator variables

$$X_{ij,1} = \begin{cases} 1 & \text{if } i = 1 \\ 0 & \text{otherwise} \end{cases}, \ldots, X_{ij,k-1} = \begin{cases} 1 & \text{if } i = k-1 \\ 0 & \text{otherwise} \end{cases}.$$

Note that if all the $k-1$ indicator variables assume the value of 0, then $i=k$. Consider the following regression model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij,1} + \cdots + \beta_{k-1} X_{ij,k-1} + \varepsilon_{ij}.$$

This model is equivalent to the factor effects model or the means model given above once we recognize that

$$\mu_1 = \beta_0 + \beta_1, \ldots, \mu_{k-1} = \beta_0 + \beta_{k-1} \text{ and } \mu_k = \beta_0.$$

Another Regression Formulation (balanced case)

Assuming a balanced case, we may define $k-1$ variables as follows

$$\begin{aligned} X_{ij,1} &= \begin{cases} 1 & \text{if } i=1 \\ -1 & \text{if } i=k \\ 0 & \text{otherwise} \end{cases}, \; X_{ij,2} = \begin{cases} 1 & \text{if } i=2 \\ -1 & \text{if } i=k \\ 0 & \text{otherwise} \end{cases}, \cdots, \\ X_{ij,k-1} &= \begin{cases} 1 & \text{if } i=1 \\ -1 & \text{if } i=k \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

Then the factor effects model can be written as

$$Y_{ij} = \mu_{\cdot} + \alpha_1 X_{ij,1} + \cdots + \alpha_{k-1} X_{ij,k-1} + \varepsilon_{ij}.$$

Note that when $i=k$,

$$\mu_{\cdot} + \alpha_1 X_{ij,1} + \cdots + \alpha_{k-1} X_{ij,k-1} = \mu_{\cdot} - \alpha_1 - \cdots - \alpha_{k-1} = \mu_{\cdot} + \alpha_k,$$

since $\sum_{i=1}^{k} \alpha_i = 0$.

For the unbalanced case, the definitions of the X-variables are slightly complicated:

$$\begin{aligned} X_{ij,1} &= \begin{cases} 1 & \text{if } i=1 \\ -n_1/n_{kr} & \text{if } i=k \\ 0 & \text{otherwise} \end{cases}, \; X_{ij,2} = \begin{cases} 1 & \text{if } i=2 \\ -n_2/n_k & \text{if } i=k \\ 0 & \text{otherwise} \end{cases}, \cdots, \\ X_{ij,k-1} &= \begin{cases} 1 & \text{if } i=1 \\ -n_{k-1}/n_k & \text{if } i=k \\ 0 & \text{otherwise} \end{cases}. \end{aligned}$$

Two-factor studies

When we have a two-factor model,

$$Y_{ij} = \mu_{ij} + \varepsilon_{ijk}, k=1,\ldots,n_{ij}, j=1,\ldots,b, i=1,\ldots,a,$$

it is customary to define

$$\begin{aligned} \mu_{\cdot\cdot} &= \frac{1}{ab}\sum\sum \mu_{ij}, \; \mu_{i\cdot} = \frac{1}{b}\sum \mu_{ij}, \; \mu_{\cdot j} = \frac{1}{a}\sum \mu_{ij}, \\ \alpha_i &= \mu_{i\cdot} - \mu_{\cdot\cdot}, \beta_j = \mu_{\cdot j} - \mu_{\cdot\cdot}, (\alpha\beta)_{ij} = \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot}, \end{aligned}$$

and hence the factor effects model is

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}.$$

The constraints now are

$$\begin{aligned}
\sum \alpha_i &= 0, \ \sum \beta_j = 0, \\
\sum_j (\alpha\beta)_{ij} &= 0 \text{ for each } i, \text{ and} \\
\sum_i (\alpha\beta)_{ij} &= 0 \text{ for each } j.
\end{aligned}$$

In order to re-express this model in the regression framework, define $a-1$ factor A variables and $b-1$ factor B variables as

$$\begin{aligned}
X^{(A)}_{ijk,1} &= \begin{cases} 1 & \text{if } i = 1 \\ -1 & \text{if } i = a \\ 0 & \text{otherwise} \end{cases}, \ldots, X^{(A)}_{ijk,a-1} = \begin{cases} 1 & \text{if } i = a-1 \\ -1 & \text{if } i = a \\ 0 & \text{otherwise} \end{cases}, \\
X^{(B)}_{ijk,1} &= \begin{cases} 1 & \text{if } j = 1 \\ -1 & \text{if } j = b \\ 0 & \text{otherwise} \end{cases}., \ldots, X^{(B)}_{ijk,b-1} = \begin{cases} 1 & \text{if } j = b-1 \\ -1 & \text{if } j = b \\ 0 & \text{otherwise} \end{cases}.
\end{aligned}$$

Then the factor effects model can be written as a regression model

$$Y_{ijk} = \mu_{..} + \sum_{l=1}^{a-1} \alpha_l X^{(A)}_{ijk,l} + \sum_{m=1}^{b-1} \beta_m X^{(B)}_{ijk,m} + \sum_{l=1}^{a-1}\sum_{m=1}^{b-1} (\alpha\beta)_{lm} X^{(A)}_{ijk,l} X^{(B)}_{ijk,m} + \varepsilon_{ijk}.$$

Note that the interaction terms $(\alpha\beta)_{lm}$ are appearing as coefficients of the product (interaction) terms of $X^{(A)}_{ijk,l}$ and $X^{(B)}_{ijk,m}$.

The additive model

$$Y_{ijk} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ijk}$$

can be re-expressed as a regression model

$$Y_{ijk} = \mu_{..} + \sum_{l=1}^{a-1} \alpha_l X^{(A)}_{ijk,l} + \sum_{m=1}^{b-1} \beta_m X^{(B)}_{ijk,m} + \varepsilon_{ijk}..$$