Math 3424 - Regression

Chapters 2 Categorical Variable, ANOVA, ANCOVA

1. One categorical variable, One-way ANOVA

1.1 Introduction

e.g., year of study: 1, 2, 3

e.g., gender: F, M

Write the model as

$$y = \beta_0 + \beta_1 x_{i1} + e_i$$

Group	Model	$E(y)$
1	$y_i = \beta_0 + \beta_1 + \text{error}$	$\beta_0 + \beta_1$
2	$y_i = \beta_0 + 2\beta_1 + \text{error}$	$\beta_0 + 2\beta_1$
3	$y_i = \beta_0 + 3\beta_1 + \text{error}$	$\beta_0 + 3\beta_1$

It is not a general model!

1.2 Regression model

Categorical variable m levels $\Rightarrow (m-1)$ dummy variables (or indicator variables)

We can use one out of three representations

Group	g_1	g_2	or	g_1	g_2	or	g_1	g_2
1	1	0		0	0		1	0
2	0	1		1	0		0	0
3	0	0		0	1		0	1

$$y = \beta_0 + \beta_{g_1} * g_1 + \beta_{g_2} * g_2 + e$$

Group	Model	$E(y)$
1	$y_i = \beta_0 + \beta_1 + \text{error}$	$\beta_0 + \beta_1 = \mu_1$
2	$y_i = \beta_0 + \beta_2 + \text{error}$	$\beta_0 + \beta_2 = \mu_2$
3	$y_i = \beta_0 + \text{error}$	$\beta_0 = \mu_3$

Model I

$$\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_{n_1} \\ \hline y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \hline y_{n_1+n_2+1} \\ \vdots \\ y_{n_1+n_2+n_3} \end{pmatrix}, \underset{\sim}{X} = \begin{pmatrix} 1 & 1 & 0 \\ \vdots & & \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \vdots & & \\ 1 & 0 & 1 \\ \hline 1 & 0 & 0 \\ \vdots & & \\ 1 & 0 & 0 \end{pmatrix}, \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \Rightarrow \hat{\underset{\sim}{\beta}},\ \hat{\sigma}^2,\ t-\text{test},\ F-\text{test}, \ldots$$

Model II

$$\underset{\sim}{Y}=\begin{pmatrix} y_1 \\ \vdots \\ y_{n_1} \\ \hline y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ \hline y_{n_1+n_2+1} \\ \vdots \\ y_{n_1+n_2+n_3} \end{pmatrix}, \underset{\sim}{X}=\begin{pmatrix} 1 & 0 & 0 \\ \vdots & & \\ 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \vdots & & \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ \vdots & & \\ 0 & 0 & 1 \end{pmatrix}, \underset{\sim}{\beta} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \Rightarrow \hat{\underset{\sim}{\beta}},\ \hat{\sigma}^2,\ t-\text{test},\ F-\text{test}, \ldots$$

$H_0 : \beta_1 = \beta_2 = 0 \Rightarrow H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_0$

1.3 ANOVA model

1.3.1 Data structure

	Factor A						
	1	2	$\cdots$	i	$\cdots$	m	
	y_{11}	y_{21}	$\cdots$	y_{i1}	$\cdots$	y_{m1}	
	y_{12}	y_{22}	$\cdots$	y_{i2}	$\cdots$	y_{m2}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	y_{1n_1}	y_{2n_2}	$\cdots$	y_{in_i}	$\cdots$	y_{mn_m}	
Total	$T_{1.}$	$T_{2.}$	$\cdots$	$T_{i.}$	$\cdots$	$T_{m.}$	$T_{..}$
Mean	$\bar{y}_{1.}$	$\bar{y}_{2.}$	$\cdots$	$\bar{y}_{i.}$	$\cdots$	$\bar{y}_{m.}$	$\bar{y}_{..}$

1.3.2 Model

1. $Y_{ij} = \mu + \alpha_i + e_{ij}$ In this model one of the α_i's is redundant. We need to place a constraint on the α_i's to avoid estimating this "extra" parameter, e.g., $\sum_{i=1}^{m} n_i\alpha_i = 0$ or $\alpha_m = 0$. For $\sum_{i=1}^{m} n_i\alpha_i = 0$,

$$\mu = \frac{\sum_{i=1}^{m} n_i\mu_i}{\sum_{i=1}^{m} n_i}$$

   For $\alpha_m = 0$, i.e., setting (m - 1) dummy variables from a categorical variable and using the last level as reference group, μ represents the mean of the observation in the last group, and $\alpha_i = \mu_i - \mu_m$.

2. Each observation may be written in the form $y_{ij} = \mu_i + e_{ij}$ where e_{ij} measures the deviation of the j^{th} observation of the i^{th} sample from the corresponding treatment mean.

   Note that the e_{ij} term represents random error and plays the same role as the error terms in the regression models. There are $m + 1$ unknown parameters: $\mu_1, \ldots, \mu_m, \sigma^2$. We estimate μ_i by the sample mean of the observations of the i^{th} group, i.e., $\mu_i = \bar{y}_{i.}$, and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{\sum_{i=1}^{m} n_i - m}.$$

1.3.3 Hypothesis testing

$H_o:\ \mu_1 = \mu_2 = \cdots = \mu_m \quad (H_0:\ \alpha_1 = \ldots = \alpha_{m-1} = \alpha_m = 0)$

$H_1:$ at least two of the means are not equal

Sum of Squares Identity

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}[(y_{ij} - \bar{y}_{..})^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}[(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + 2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})(y_{ij} - \bar{y}_{i.}) \\
&\quad + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2
\end{aligned}$$

The middle term is zero. Hence

$$\begin{aligned}
\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \\
&= \sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \\
&= n\sum_{i=1}^{m}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{i.})^2
\end{aligned}$$

if $n_i = n$ for all i.

Symbolically, we write the sum of squares identity as

$$SST \ = \ SSA + SSE$$

where SST is called total sum of squares, SSA is called treatment sum of squares and SSE is the error sum of squares. The degrees of freedom are partitioned according to the identity

$$\sum_{i=1}^{m} n_i - 1 \ = \ (m-1) + \sum_{i=1}^{m} n_i - m$$

It can be shown that

$$\begin{aligned}
E\left(\frac{SSA}{m-1}\right) &= \sigma^2 + \frac{\sum_{i=1}^{m} n_i\alpha_i^2}{m-1} \\
E\left(\frac{SSE}{\sum_{i=1}^{m} n_i - m}\right) &= \sigma^2
\end{aligned}$$

Source of Variation of	Sum of Squares	Degrees of freedom	Mean Square	Computed f
Model	$\sum_{i=1}^{m} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	$m-1$	$\frac{\sum_{i=1}^{m} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{m-1}$	$\frac{(\sum_{i=1}^{m} n_i - m)\sum_{i=1}^{m} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{(m-1)\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2}$
Error	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	$\sum_{i=1}^{m} n_i - m$	$\frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2}{\sum_{i=1}^{m} n_i - m}$	
Total	$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$	$\sum_{i=1}^{m} n_i - 1$		

The advantages of choosing equal sample sizes over the choice of unequal sample sizes are: 1) the f ratio is insensitive to slight departures from the assumption of equal variances for the m populations when the sample are of equal sizes; and 2) the choice of equal sample size minimizes the probability of committing a type II error.

Example

	Group					
	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16854
Mean	553.33	569.33	610.50	465.19	610.67	561.80

Solution

$S_1^2 = 12,133.8667$, $S_2^2 = 2,302.6667$, $S_3^2 = 3593.5$, $S_4^2 = 3,318.5667$, $S_5^2 = 3,455.4667$, $S_T^2 = 7,219.8897$

Source of Variation of	Sum of Squares	Degrees of freedom	Mean Square	Computed f
Group	85,356	4	21,339	4.30
Error	124,021	25	4,961	
Total	209,377	29		

The critical value $f_{0.05}(4, 25) = 2.76$. Thus, $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ is rejected.

1.3.4 Single-degree-of-freedom Comparisons

Definition Any linear function of the form

$$\omega = \sum_{i=1}^{m} c_i \mu_i, \quad \text{where } \sum_{i=1}^{m} c_i = 0,$$

is called a comparison or contrast in the treatment means.

The experimenter can often make multiple comparisons by testing the significance of contrasts in the treatment means, that is, by testing a hypothesis of the type

$$H_0 : \sum_{i=1}^{m} c_i \mu_i = 0,$$

$$H_1 : \sum_{i=1}^{m} c_i \mu_i \neq 0,$$

where $\sum_{i=1}^{m} c_i = 0$. The test is conducted by first computing a similar contrast in the sample means,

$$\hat{\omega} = \sum_{i=1}^{m} c_i \bar{y}_{i.} \ .$$

$\hat{\omega}$ is a value of the normal random variable ω with mean

$$\mu_\omega \ = \sum_{i=1}^{m} c_i \mu_i$$

and variance

$$\sigma_\omega^2 \ = \ \sigma^2 \sum_{i=1}^{m} \frac{c_i^2}{n_i} \ .$$

Therefor, when H_0 is true, $\mu_\omega = 0$ and the statistic

$$\frac{\left(\sum_{i=1}^{m} c_i \bar{Y}_{i.}\right)^2}{\sigma^2 \sum_{i=1}^{m} (c_i^2/n_i)}$$

is distributed as a chi-square random variable with 1 degree of freedom. Our hypothesis is tested at the α level of significance by computing

$$f \ = \ \frac{\left(\sum_{i=1}^{m} c_i \bar{Y}_{i.}\right)^2}{\hat{\sigma}^2 \sum_{i=1}^{m} (c_i^2/n_i)} \ = \ \frac{\left[\sum_{i=1}^{m} (c_i T_{i.}/n_i)\right]^2}{\hat{\sigma}^2 \sum_{i=1}^{m} (c_i^2/n_i)} \ = \ \frac{SSW}{\hat{\sigma}^2}$$

where f is a value of the random variable F having the F distribution with 1 and $\sum_{i=1}^{m} n_i - m$ degrees of freedom and

$$SSW \ = \ \frac{\left[\sum_{i=1}^{m} (c_i T_{i.}/n_i)\right]^2}{\sum_{i=1}^{m} (c_i^2/n_i)} \ .$$

When the sample sizes are all equal to n,

$$SSW \ = \ \frac{\left(\sum_{i=1}^{m} c_i T_{i.}\right)^2}{n \sum_{i=1}^{m} c_i^2} \ .$$

Definition The two contrasts

$$\omega_1 \ = \ \sum_{i=1}^{m} c_i \mu_i \ \text{ and } \ \omega_2 \ = \ \sum_{i=1}^{m} d_i \mu_i$$

are said to be orthogonal if $\sum_{i=1}^{m} c_i d_i / n_i \ = \ 0$ or when the n_i's are all equal to n if $\sum_{i=1}^{m} c_i d_i = 0$.

If ω_1 and ω_2 are orthogonal, then the quantities SSW_1 and SSW_2 are components of SSA (i.e., S.S. for group in our example), each with a single degree of freedom. The treatment sum of squares with $m-1$ degrees of freedom can be partitioned into at most $m-1$ independent single-degree-of-freedom contrast sum of squares satisfying the identity

$$SSA = SSW_1 + SSW_2 + \ldots + SSW_{m-1}$$

if the contrasts are orthogonal to each other.

Example

Find the contrast sum of squares corresponding to the orthogonal contrasts

$$\omega_1 = \mu_1 + \mu_2 + \mu_3 + \mu_5 - 4\mu_4$$

$$\omega_2 = \mu_1 + \mu_2 - \mu_3 - \mu_5$$

and carry out appropriate tests of significance.

Solution

One can write down two additional contrasts orthogonal to the first two such as

$$\omega_3 = \mu_1 - \mu_2$$

$$\omega_4 = \mu_3 - \mu_5$$

Source of Variation of	Sum of Squares	Degrees of freedom	Mean Square	Computed f
Groups	85,356	4	21,339	4.30
(1,2,3,5) vs 4	70,035	1	70,035	14.12
(1,2) vs (3,5)	14,553	1	14,553	2.93
(1) vs (2)	768	1	768	0.15
(3) vs (5)	0.08	1	0.08	0.00
Error	124,021	25	4,961	
Total	209,377	29		

The contrast ω_1 is not significant when compared to the critical value $f_{0.05}(1, 25) = 4.24$. However, the f value of 14.12 for ω_2 is significant and the hypothesis

$$H_0 : \mu_1 + \mu_2 + \mu_3 + \mu_5 = 4\mu_4$$

is rejected.

2. Two categorical variables, Two-way ANOVA

2.1 Regression model

group	–	3 levels	$1, 2, 3$	$\Rightarrow$	g_1, g_2
class	–	3 levels	A, B, C	$\Rightarrow$	c_1, c_2

Group	g_1	g_2
1	1	0
2	0	1
3	0	0

Class	c_1	c_2
A	1	0
B	0	1
C	0	0

Model I

$$y = \beta_0 + \beta_{g_1} * g_1 + \beta_{g_2} * g_2 + \beta_{c_1} * c_1 + \beta_{c_2} * c_2 + \beta_{g_1,c_1}(g_1 * c_1) + \beta_{g_1,c_2}(g_1 * c_2) + \beta_{g_2,c_1}(g_2 * c_1) + \beta_{g_2,c_2}(g_2 * c_2) + e$$

Model II

$y_{ijk} = \mu_{ij} + e$ for $i, j = 1, 2, 3$; $k = 1, \ldots, n_{ij}$

$H_0 : \beta_{g_1,c_1} = \beta_{g_1,c_2} = \beta_{g_2,c_1} = \beta_{g_2,c_2} = 0$ (no interaction)
$\Rightarrow H_0 : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}, \mu_{21} - \mu_{31} = \mu_{22} - \mu_{32} = \mu_{23} - \mu_{33}$ (parallel curves)

2.2 ANOVA model

2.2.1 Data structure

Denote the k^{th} observation taken at the i^{th} level of factor A and the $j^(th)$ level of factor B by y_{ijk}. For simplicity, here we focus on the cases with balanced data, i.e., the observations in the $(i, j)^{th}$ cell constitute a sample with the same sample size n. The abn observations are shown as follows.

	Factor B				
Factor A	1	2	$\cdots$	b	Mean
1	y_{111}	y_{121}	$\cdots$	y_{1b1}	$\bar{y}_{1..}$
	y_{112}	y_{122}	$\cdots$	y_{1b2}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{11n}	y_{12n}	$\cdots$	y_{1bn}	
2	y_{211}	y_{221}	$\cdots$	y_{2b1}	$\bar{y}_{2..}$
	y_{212}	y_{222}	$\cdots$	y_{2b2}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{21n}	y_{22n}	$\cdots$	y_{2bn}	
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
a	y_{a11}	y_{a21}	$\cdots$	y_{ab1}	$\bar{y}_{a..}$
	y_{a12}	y_{a22}	$\cdots$	y_{ab2}	
	$\vdots$	$\vdots$	$\cdots$	$\vdots$	
	y_{a1n}	y_{a2n}	$\cdots$	y_{abn}	
Mean	$\bar{y}_{.1.}$	$\bar{y}_{.2.}$	$\cdots$	$\bar{y}_{.b.}$	$\bar{y}_{...}$

2.2.2 Model

1.

$$\begin{aligned} y_{ijk} &= \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk} \quad i = 1, 2, \ldots, a,\ j = 1, 2, \ldots, b,\ k = 1, 2, \ldots, n \\ e_{ijk} &\sim N(0, \sigma^2) \quad \text{and independent} \end{aligned}$$

Constraints:

(a) $\sum_{i=1}^{a} \alpha_i = 0,\ \sum_{i=1}^{b} \beta_i = 0,\ \sum_{i=1}^{a} \gamma_{ij} = 0,\ \sum_{j=1}^{b} \gamma_{ij} = 0$; or

(b) $\alpha_a = 0,\ \beta_b = 0,\ \gamma_{aj} = 0$ for each j, $\gamma_{ib} = 0$ for each i
In this case, μ represents the mean of the observation in the last combination, i.e., $\mu = \mu_{ab}$

2. The two-way ANOVA can be written as

$$\begin{aligned} y_{ijk} &= \mu_{ij} + e_{ijk} \quad i = 1, 2, \ldots, a,\ j = 1, 2, \ldots, b,\ k = 1, 2, \ldots, n \\ e_{ijk} &\sim N(0, \sigma^2) \quad \text{and independent} \end{aligned}$$

2.2.3 Hypothesis testing

The three hypotheses to be tested are as follows:

1. $H_0^1 : \alpha_1 = \alpha_2 = \ldots = \alpha_a = 0$
2. $H_0^2 : \beta_1 = \beta_2 = \ldots = \beta_b = 0$
3. $H_0^3 : \gamma_{11} = \gamma_{12} = \ldots = \gamma_{ab} = 0$

Or

1. No Difference in Means Due to Factor A
   $H_0^1 : \mu_{1.} = \mu_{2.} = \ldots = \mu_{a.}$
2. No Difference in Means Due to Factor B
   $H_0^2 : \mu_{.1} = \mu_{.2} = \ldots = \mu_{.b}$
3. No Interaction of Factors A & B

$$\begin{aligned} H_0^3 : \quad & \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \ldots = \mu_{1b} - \mu_{2b}; \\ & \mu_{11} - \mu_{31} = \mu_{12} - \mu_{32} = \ldots = \mu_{1b} - \mu_{3b}; \\ & \quad \vdots \\ & \mu_{11} - \mu_{a1} = \mu_{12} - \mu_{a2} = \ldots = \mu_{1b} - \mu_{ab} \end{aligned}$$

Analysis

1. Consider a model with interaction.
2. Test "interaction" effect.
3. If the interaction terms are not significant, test "Factor A" effect & test "Factor B" effect (main effect comparison).
4. if the interaction is significant, only simple effect comparison is appropriate, i.e., if we want to study the effect of one factor, e.g., "method", we should look separately at three levels of variety, i.e., variety 1, variety 2 & variety 3.
5. In the presence of significant interaction term, the main effect components should be included for the ease of interpretation even they are insignificant.

2.2.4 Balanced design

Sum of Squares Identity

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}[(\bar{y}_{i..}-\bar{y}_{...})+(\bar{y}_{.j.}-\bar{y}_{...})+ \\
&\quad (\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})+(y_{ijk}-\bar{y}_{ij.})]^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{i..}-\bar{y}_{...})^2+\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{.j.}-\bar{y}_{...})^2 \\
&\quad +\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 \\
&\quad +\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2 \\
&\quad +6\ \text{cross}-\text{product terms}
\end{aligned}$$

The cross-product terms are all equal to zero for balanced design. Hence

$$\begin{aligned}
\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{...})^2 &= bn\sum_{i=1}^{a}(\bar{y}_{i..}-\bar{y}_{...})^2+an\sum_{j=1}^{b}(\bar{y}_{.j.}-\bar{y}_{...})^2 \\
&\quad +n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..}-\bar{y}_{.j.}+\bar{y}_{...})^2 \\
&\quad +\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk}-\bar{y}_{ij.})^2
\end{aligned}$$

Symbolically, we write the sum of squares identity as

$$SST = SS(A)+SS(B)+SS(AB)+SSE$$

where $SS(A)$ and $SS(B)$ are called the sum of squares for the main effects A and B respectively, $SS(AB)$ is called the interaction sum of squares for A and B, and SSE is the error sum of squares. The degrees of freedom are partitioned according to the identity

$$abn-1 = (a-1)+(b-1)+(a-1)(b-1)+ab(n-1)$$

It can be shown that

$$\begin{aligned}
E\left(\frac{SS(A)}{a-1}\right) &= \sigma^2+\frac{nb\sum_{i=1}^{a}\alpha_i^2}{a-1} \\
E\left(\frac{SS(B)}{b-1}\right) &= \sigma^2+\frac{na\sum_{j=1}^{b}\beta_j^2}{b-1} \\
E\left(\frac{SS(AB)}{(a-1)(b-1)}\right) &= \sigma^2+\frac{n\sum_{i=1}^{a}\sum_{j=1}^{b}\gamma_{ij}^2}{(a-1)(b-1)} \\
E\left(\frac{SSE}{ab(n-1)}\right) &= \sigma^2
\end{aligned}$$

Source of Variation of	Sum of Squares	Degrees of freedom	Mean Square	Computed f
A	SSA	$a-1$	$s_1^2 = \frac{SSA}{a-1}$	
B	SSB	$b-1$	$s_2^2 = \frac{SSB}{b-1}$	
AB	$SS(AB)$	$(a-1)(b-1)$	$s_3^2 = \frac{SS(AB)}{(a-1)(b-1)}$	$f_3 = \frac{s_3^2}{s^2}$
Error	SSE	$ab(n-1)$	$s^2 = \frac{SSE}{ab(n-1)}$	
Total	$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{...})^2$	$abn-1$		

$SS(A) = bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2$, $SS(B) = an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$,

$SS(AB) = n\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$, $SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2$

Remarks

1. We should first observe whether or not the test for interaction is significant.

2. If the interaction is insignificant, then the results of the tests on the main effects are meaningful, but most often we will redo the two-way ANOVA model again without the interaction term.

   This will put the SS and df for $A * B$ into Error.

   Results of main effect hypothesis tests could change because MSE and denominator df have changed (more impact with small sample size). Then,

$$f_1 = \frac{s_1^2}{\hat{\sigma}^2_{\text{no int}}} \quad \text{for Factor A}$$

$$f_2 = \frac{s_2^2}{\hat{\sigma}^2_{\text{no int}}} \quad \text{for Factor B}$$

   where

$$\hat{\sigma}^2_{\text{no int}} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \bar{y}_{ij.})^2 \; + \; SS(AB)}{ab(n-1) + (a-1)*(b-1)}$$

Example

	Variety				
Method	1	2	3	Sum	CSS
1	22.3	19.8	20		
	25.8	28.3	17		
	22.8	26.8	24		
	28.3	27.3	22.5		
	21.3	26.8	28		
	18.3	26.8	22.5		
Sum	138.8	155.8	134	428.6	
Corrected S.S.	61.333333	47.333333	68.833333	221.237778	
2	16.4	24.5	11.8		
	14.4	16	14.3		
	21.4	11	21.3		
	19.9	7.5	6.3		
	10.4	14.5	7.8		
	21.4	15.5	13.8		
Sum	103.9	89	75.3	268.2	
Corrected S.S.	97.208333	163.833333	143.375	472.62	
Sum	242.7	244.8	209.3	696.8	
Corrected S.S.	260.0425	583.02	499.349167	1408.53	

Source of Variation of	Sum of Squares	Degrees of freedom	Mean Square	Computed f
Method	714.671111	1	714.671111	36.84
Variety	66.117222	2	33.058611	1.71
Interaction	45.823889	2	22.911944	1.18
Error	581.916667	30	19.397222	
Total	1408.528889	35		

Test "interaction" effect is equivalent to test $H_0 : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$.

As the interaction terms are not significant, we re-construct the ANOVA table.

Source of Variation of	Sum of Squares	Degrees of freedom	Mean Square	Computed f
Method	714.671111	1	714.671111	36.43
Variety	66.117222	2	33.058611	1.69
Error	627.740556	32	19.616892	
Total	1408.528889	35		

Test "method" effect is equivalent to test $H_0 : \mu_{11} + \mu_{12} + \mu_{13} = \mu_{21} + \mu_{22} + \mu_{23}$, i.e., $H_0 : \mu_{1.} = \mu_{2.}$

Test "variety" effect is equivalent to test $H_0 : \mu_{11} + \mu_{21} = \mu_{12} + \mu_{22} = \mu_{13} + \mu_{23}$, i.e., $H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3}$

Remarks

1. SS(A) & SS(B) can be calculated from the expression in the "sum of squares" identity, i.e.,

$$\begin{aligned} \text{SS(A)} &= bn\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2 \\ \text{SS(B)} &= an\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2 \end{aligned}$$

2. SS(A) & SS(B) for the model without interaction are equal to those for the model with interaction.
3. "error" S.S. for the model without interaction is equal to the sum of "interaction" S.S. & "error" S.S. for the model with interaction.

2.2.5 Unbalanced design

1. • In "sum of squares identity"
   (a) The cross-product terms are not equal to zero.
   (b) $SS(A) = R(A|\beta_0)$ & $SS(B) = R(B|\beta_0)$
   • In the regression model, $SS(A) = R(A|B, AB, \beta_0)$ & $SS(B) = R(B|A, AB, \beta_0)$ – partial S.S.
2. Sum of squares is calculated as $(\underset{\sim}{C}\hat{\underset{\sim}{\beta}})^T[\underset{\sim}{C}(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\underset{\sim}{C}^T]^{-1}(\underset{\sim}{C}\hat{\underset{\sim}{\beta}})$ by choosing an appropriate $\underset{\sim}{C}$.
3. For the model with interaction term(s)
   •
   $$\begin{aligned} \text{SSE} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij.})^2 \\ \text{d.f.} &= \sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij} - 1) \end{aligned}$$
   • Test "interaction" effect is equivalent to test $H_0 : \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23}$. Then,
   $$\underset{\sim}{C} = \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{pmatrix} \quad \hat{\underset{\sim}{\beta}} = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{pmatrix}$$
4. If the interaction terms are not significant,
   •
   $$\begin{aligned} \text{SSE} &= \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij.})^2 \ + \ SS(AB) \\ \text{d.f.} &= \sum_{i=1}^{a}\sum_{j=1}^{b}(n_{ij} - 1) + (a-1)*(b-1) \end{aligned}$$
   • Test "method" effect is equivalent to test $H_0 : \mu_{1.} = \mu_{2.} \ \Rightarrow \ H_0 : \alpha_1 = 0$. Then,
   $$\underset{\sim}{C} = \begin{pmatrix} 0 & 1 & 0 & 0 \end{pmatrix} \quad \hat{\underset{\sim}{\beta}} = \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$
   • Test "variety" effect is equivalent to test $H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} \ \Rightarrow \ H_0 : \beta_1 = \beta_2 = 0$. Then,
   $$\underset{\sim}{C} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \quad \hat{\underset{\sim}{\beta}} = \begin{pmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

5. With unbalanced designs, sample mean will give us the averages of all observations in each level of a factor, which is the unadjusted ("biased") cell means. Using the estimates of cell means from the model can produce adjusted (for the contamination effects of other factors in the model) ("unbiased") means.

   - $$\begin{aligned} \bar{y}_{1..} &= \frac{\sum_{k=1}^{n_{11}} y_{11k} + \sum_{k=1}^{n_{12}} y_{12k}}{n_{11}+n_{12}} \\ E(\bar{y}_{1..}) &= \mu + \alpha_1 + \frac{n_{11}}{n_{11}+n_{12}}\beta_1 + \frac{n_{12}}{n_{11}+n_{12}}\beta_2 \\ \text{But} \quad \mu_{1..} &= \mu + \alpha_1 + \frac{1}{2}\beta_1 + \frac{1}{2}\beta_2 \end{aligned}$$

   - $$\begin{aligned} \bar{y}_{1..} - \bar{y}_{2..} &= \frac{\sum_{k=1}^{n_{11}} y_{11k} + \sum_{k=1}^{n_{12}} y_{12k}}{n_{11}+n_{12}} - \frac{\sum_{k=1}^{n_{21}} y_{21k} + \sum_{k=1}^{n_{22}} y_{22k}}{n_{21}+n_{22}} \\ E(\bar{y}_{1..} - \bar{y}_{2..}) &= \alpha_1 - \alpha_2 + \left(\frac{n_{11}}{n_{11}+n_{12}} - \frac{n_{21}}{n_{21}+n_{22}}\right)\beta_1 + \left(\frac{n_{11}}{n_{11}+n_{12}} - \frac{n_{22}}{n_{21}+n_{22}}\right)\beta_2 \end{aligned}$$

     The difference between Factor A effect is biased by Factor B effect.

3. ANCOVA

3.1 One categorical variable + one cont. variable

Table 3.3 Data involving three sets of subjects

	Group	Weight (lb)	HDL Cholesterol (mg/decaliter)
Control	1	163.5	75
	1	180	72.5
	1	178.5	62
	1	161.5	60
	1	127	53
	1	161	53
	1	165	65
	1	144	63.5
Running	2	141	49
	2	162	53.5
	2	134	30
	2	121	40.5
	2	145	51.5
	2	106	57.5
	2	134	49
	2	216.5	74
Running and Weightlifting	3	136.5	54.5
	3	142.5	79.5
	3	145	64
	3	165	69
	3	226	50.5
	3	122	58
	3	193	63.5
	3	163.5	76
	3	154	55.5
	3	139	68

Model I

$$y = \beta_0 + \beta_{g_1} * g_1 + \beta_{g_2} * g_2 + \beta_1 * x + \beta_{1g_1} * g_1 * x + \beta_{1g_2} * g_2 * x + e$$

Model II

$$\begin{aligned} A &: y_i = \gamma_{01} + \gamma_{11} x_i + e_i, && i = 1, \ldots, 8 \\ B &: y_i = \gamma_{02} + \gamma_{12} x_i + e_i, && i = 9, \ldots, 16 \\ C &: y_i = \gamma_{03} + \gamma_{13} x_i + e_i, && i = 17, \ldots, 26 \end{aligned}$$

$$\underset{\sim}{Y} = \begin{pmatrix} 75 \\ 72.5 \\ \vdots \\ 68 \end{pmatrix} \qquad \underset{\sim}{\beta} = \begin{pmatrix} \gamma_{01} \\ \gamma_{11} \\ \gamma_{02} \\ \gamma_{12} \\ \gamma_{03} \\ \gamma_{13} \end{pmatrix}$$

$$\underset{\sim}{X} = \begin{pmatrix} 1 & 163.5 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 144 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 141 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 216.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 136.5 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & 139 \end{pmatrix}$$

Control

$$\mathrm{E}(y) = \beta_0 + \beta_{g_1} + (\beta_1 + \beta_{1g_1})\text{weight} = \gamma_{01} + \gamma_{11}\ \text{weight}$$

Running

$$\mathrm{E}(y) = \beta_0 + \beta_{g_2} + (\beta_1 + \beta_{1g_2})\text{weight} = \gamma_{02} + \gamma_{12}\ \text{weight}$$

Running and Weighting

$$\mathrm{E}(y) = \beta_0 + \beta_1\text{weight} = \gamma_{03} + \gamma_{13}\ \text{weight}$$

$$\hat{\beta}_0 = 76.88002 \qquad \hat{\beta}_{g1} = 23.5431 - 76.88002 = -53.82571 \qquad \hat{\beta}_{g2} = -62.62502$$
$$\hat{\beta}_1 = -0.08213 \qquad \hat{\beta}_{1g_1} = 0.24956 - (-0.08213) = 0.33169 \qquad \hat{\beta}_{1g_2} = 0.33307$$

$\Rightarrow y = 76.88 - 53.83g_1 - 62.63g_2 - 0.08213\text{weight} + 0.33169(g_1 * \text{weight}) + 0.033307(g_2 * \text{ weight}) + e$

- Estimate of σ^2

- F-test

- $H_0 : \beta_{1g1} = \beta_{1g2} = 0$ (no interaction) $\Rightarrow H_0 : \gamma_{11} = \gamma_{12} = \gamma_{13}$ (parallel lines)

3.2 Two categorical variables + one cont. variable

Model I

$$\begin{aligned} y = {} & \beta_0 + \beta_{g_1} * g_1 + \beta_{g_2} * g_2 + \beta_{c_1} * c_1 + \beta_{c_2} * c_2 + \beta_{g_1,c_1}(g_1 * c_1) + \\ & \beta_{g_1,c_2}(g_1 * c_2) + \beta_{g_2,c_1}(g_2 * c_1) + \beta_{g_2,c_2}(g_2 * c_2) + \beta_1 * x + \beta_{g_1,x}(g_1 * x) + \beta_{g_2,x}(g_2 * x) + \\ & \beta_{c_1,x}(c_1 * x) + \beta_{c_2,x}(c_1 * x) + \beta_{g_1,c_1,x}(g_1 * c_1 * x) + \\ & \beta_{g_1,c_2,x}(g_1 * c_2 * x) + \beta_{g_2,c_1,x}(g_2 * c_1 * x) + \beta_{g_2,c_2,x}(g_2 * c_2 * x) + e \end{aligned}$$

Model II

$y_{ijk} = \gamma_{0ij} + \gamma_{1ij} + e$ for $i, j = 1, 2, 3$; $k = 1, \ldots, n_{ij}$

$H_0 : \beta_{g_1,x} = \beta_{g_2,x} = \beta_{c_1,x} = \beta_{c_2,x} = \beta_{g_1,c_1,x} = \beta_{g_1,c_2,x} = \beta_{g_2,c_1,x} = \beta_{g_2,c_2,x} = 0$ (no interaction)
$\Rightarrow H_0 : \gamma_{111} = \gamma_{112} = \gamma_{113} = \gamma_{121} = \gamma_{122} = \gamma_{123} = \gamma_{131} = \gamma_{132} = \gamma_{133}$ (parallel curves)