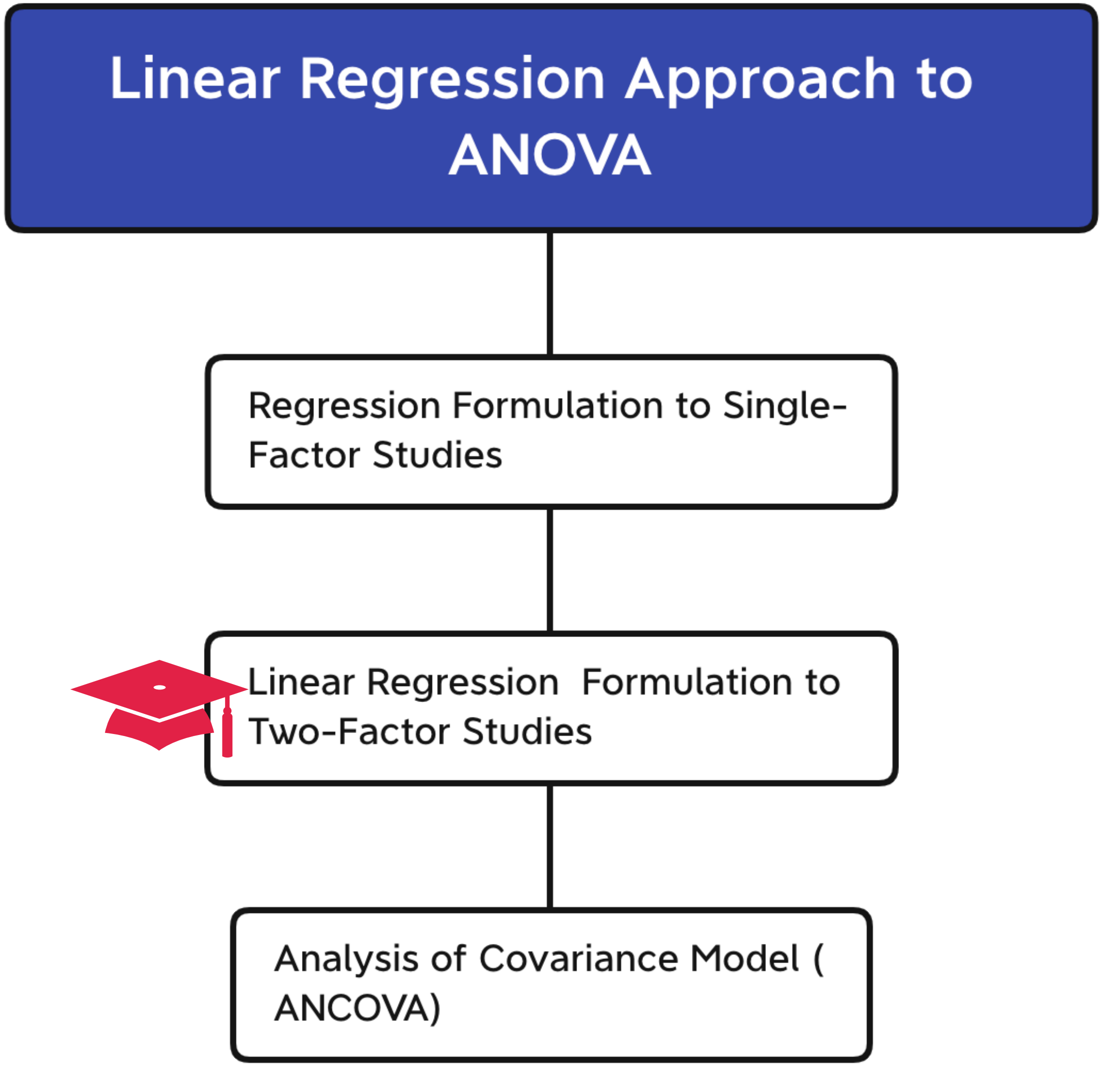
Lecture 6: Linear Regression Approach to ANOVA

STA 106: Analysis of Variance

Xiner Zhou
Department of Statistics, University of California, Davis


Linear Regression Approach to ANOVA
Regression Formulation to Single-Factor Studies
Linear Regression Formulation to Two-Factor Studies
Analysis of Covariance Model (ANCOVA)

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

Why unequal sample sizes?

- Observational studies
  - investigators usually has no control over how many sample sizes for each treatment or cell, the sample sizes are determined by what data can be collected and be available for analysis
- Experimental studies
  - even under the design of equal sample size, the data often end up having unequal sample sizes
    - illness of subjects
    - lost to follow up
    - technical problems
    - ......
  - by design

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

Analysis of Variance

Partition of Total Sum of Squares

$$Y_{ijk} - \bar{Y}_{\dots} = \bar{Y}_{ij\cdot} - \bar{Y}_{\dots} + Y_{ijk} - \bar{Y}_{ij\cdot}$$

Total deviation

Deviation of estimated treatment mean around overall mean

Deviation around estimated treatment mean

$$\underbrace{\sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{\dots}\right)^2}_{\text{Total variation}} = \sum_i \sum_j \sum_k \left(\bar{Y}_{ij\cdot} - \bar{Y}_{\dots}\right)^2 + \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2$$

$$= \underbrace{n \sum_i \sum_j \left(\bar{Y}_{ij\cdot} - \bar{Y}_{\dots}\right)^2}_{\text{Variation due to factor A and B}} + \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2$$

Let $SSTO = \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{\dots}\right)^2$

$$SSTR = n \sum_i \sum_j \left(\bar{Y}_{ij} - \bar{Y}_{\dots}\right)^2$$

$$SSE = \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{ij\cdot}\right)^2 = \sum_i \sum_j \sum_k e_{ijk}^2$$

$$\boxed{SSTO \quad = \quad SSTR \quad + \quad SSE}$$

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

Analysis of Variance

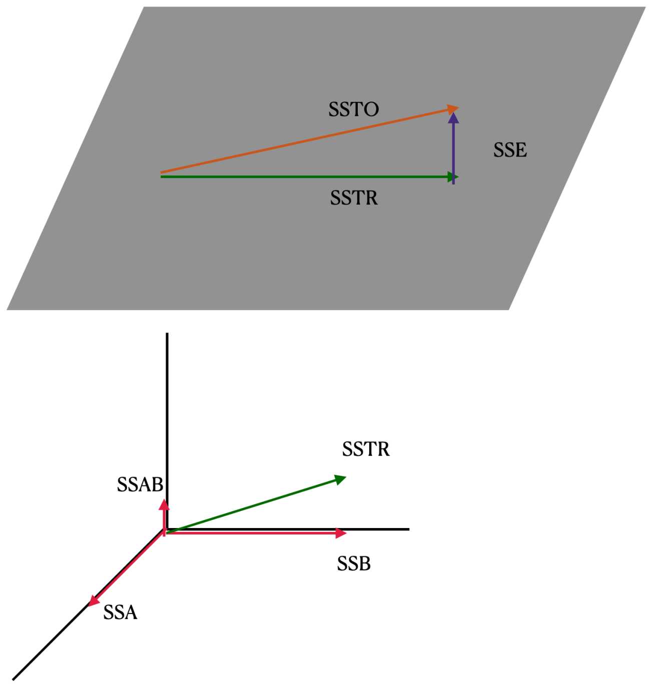
Partition of Treatment Sum of Squares.

$$\underbrace{\bar{Y}_{ij} - \bar{Y}_{\dots}}_{\text{Deviation of estimated treatment mean around overall mean}} = \underbrace{\bar{Y}_{i..} - \bar{Y}_{...}}_{\text{A main effect}} + \underbrace{\bar{Y}_{.j.} - \bar{Y}_{...}}_{\text{B main effect}} + \underbrace{\bar{Y}_{ij} - \bar{Y}_{i..} - \bar{Y}_{j.} + \bar{Y}_{...}}_{\text{AB interaction effect}}$$

$$\underbrace{\sum_i \sum_j \sum_k}_{\text{SSTR : treatment sum of squares}} \left(\bar{Y}_{ij} - \bar{Y}_{\dots}\right)^2 = \underbrace{bn \sum_i \left(\bar{Y}_{i..} - \bar{Y}_{...}\right)^2}_{\text{SSA: factor A sum of squares}} + \underbrace{an \sum_j \left(\bar{Y}_{.j.} - \bar{Y}_{...}\right)^2}_{\text{SSB: factor B sum of squares}} + \underbrace{\sum_i \sum_j \sum_k \left(\bar{Y}_{ij} - \bar{Y}_{i..} - \bar{Y}_{j.} + \bar{Y}_{...}\right)^2}_{\text{AB interaction sum of squares}}$$

SSTR = SSA + SSB +SSAS

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes



Loss of the nice decomposition of variance

—> can't disentangle different forces (factor A, factor B) into the system

—> More general way to decompose "intertwined" signals

—> linear regression approach

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

Two-Way ANOVA Model

$$Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \qquad \text{Treatment means parameterization}$$

$$= \mu_{..} + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad \text{Factor effects parameterization}$$

- $\mu_{..} = \dfrac{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}}{ab}$: overall mean
- $\alpha_i = \mu_{i.} - \mu_{..}$ main effect of factor A at ith level

  Subject to (a-1) constraints $\sum \alpha_i = 0$
- $\beta_j = \mu_{.j} - \mu_{..}$ main effect of factor B at jth level

  Subject to (b-1) constraints $\sum \beta_j = 0$
- $\gamma_{ij} = \mu_{ij} - \alpha_i - \beta_j = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$ interaction effect of factor A at ith level with factor B at jth level

  Subject to a+b-1 constraints

  $$\sum_i (\alpha\beta)_{ij} = 0 \;\; j = 1,\ldots,b$$
  $$\sum_j (\alpha\beta)_{ij} = 0 \;\; i = 1,\ldots,a$$
- ε_{ijk} are independent $N\left(0,\sigma^2\right)$ for $i = 1,\ldots,a; j = 1,\ldots,b; k = 1,\ldots,n_{ij}$

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

Indicator variables for factor A

$$X_{ijk;1,.} = \begin{cases} 1 \text{ if case from level } 1 \text{ for factor } A \\ -1 \text{ if case from level } a \text{ for factor } A \\ 0 \text{ otherwise} \end{cases}$$

$\ldots$

$$X_{ijk;a-1,.} = \begin{cases} 1 \text{ if case from level } a-1 \text{ for factor } A \\ -1 \text{ if case from level } a \text{ for factor } A \\ 0 \text{ otherwise} \end{cases}$$

Indicator variables for factor B

$$X_{ijk;.,1} = \begin{cases} 1 \text{ if case from level } 1 \text{ for factor } B \\ -1 \text{ if case from level } a \text{ for factor } B \\ 0 \text{ otherwise} \end{cases}$$

$\ldots$

$$X_{ijk;.,b-1} = \begin{cases} 1 \text{ if case from level } b-1 \text{ for factor } B \\ -1 \text{ if case from level } a \text{ for factor } B \\ 0 \text{ otherwise} \end{cases}$$

$$Y_{ijk} = \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \ldots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk;.,1} + \ldots + \beta_{b-1} X_{ijk;.,b-1}}_{\text{B main effect}}$$

$$+ \underbrace{\gamma_{11} X_{ijk,1,.} X_{ijk,.,1} + \ldots + \gamma_{(a-1)(b-1)} X_{ijk,a-1,.} X_{ijk,.,b-1}}_{\text{AB interaction effect}} + \varepsilon_{ijk}$$

Regression coefficients $\mu_{..}, \alpha_1, \ldots, \alpha_{a-1}, \beta_1, \ldots, \beta_{b-1}, \gamma_{11}, \ldots, \gamma_{(a-1)(b-1)}$ estimated by least squares method

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

In One-Way ANOVA and Two-Way ANOVA with equal sample size, it reduce to the usual F tests

because of the nice partition of sum of squares SSTO=SSA+SSB+SSAB+SSE

and the nice geometry

allow to disentangle variations due to different sources easily

But beyond above two simple situations, such as in Two-Way ANOVA with unequal sample sizes or multi-factor studies with unequal sample sizes

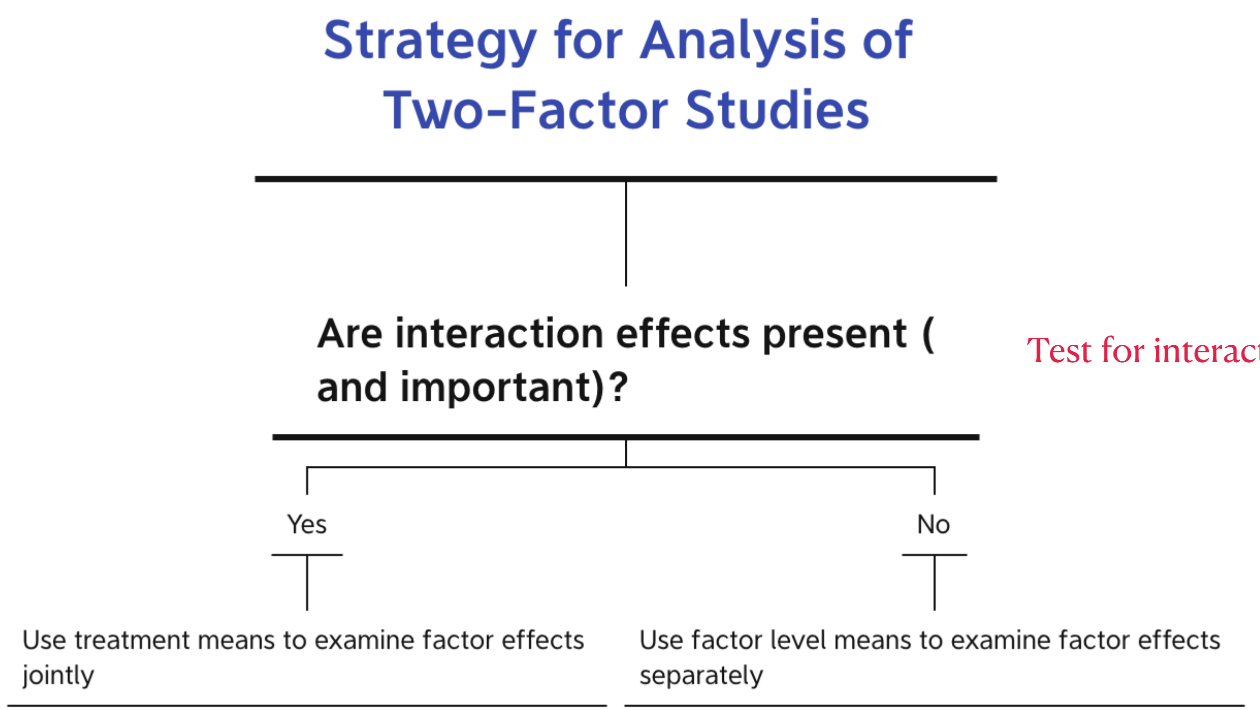
the nice partition of sum of squares SSTO=SSA+SSB+SSAB+SSE and the nice geometry no longer hold

So we need more general analysis of variance approach in the regression analysis :

General Linear Test Approach: test about regression parameters

Idea: whether there is significant reduction in the error variance (measured by SSE and MSE) when a variable or a set of variables added to the regression model

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes



Test for interaction effects

Test for main effects

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

Test for Interaction Effects

$H_0 : \gamma_{11} = \ldots = \gamma_{(a-1)(b-1)} = 0$ $\quad$ $H_a :$ not all $\gamma_{11} \ldots \gamma_{(a-1)(b-1)}$ equal zero

Full model

$$Y_{ijk} = \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \ldots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk;.,1} + \ldots + \beta_{b-1} X_{ijk;.,b-1}}_{\text{B main effect}}$$

$$+ \underbrace{\gamma_{11} X_{ijk,1,.} X_{ijk,.,1} + \ldots + \gamma_{(a-1)(b-1)} X_{ijk,a-1,.} X_{ijk,.,b-1}}_{\text{AB interaction effect}} + \varepsilon_{ijk}$$

Reduced model

$$Y_{ijk} = \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \ldots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk;.,1} + \ldots + \beta_{b-1} X_{ijk;.,b-1}}_{\text{B main effect}}$$

F test:

$$F^* = \frac{\text{SSE}(R) - \text{SSE}(F)}{df_R - df_F} \div \frac{\text{SSE}(F)}{df_F}$$

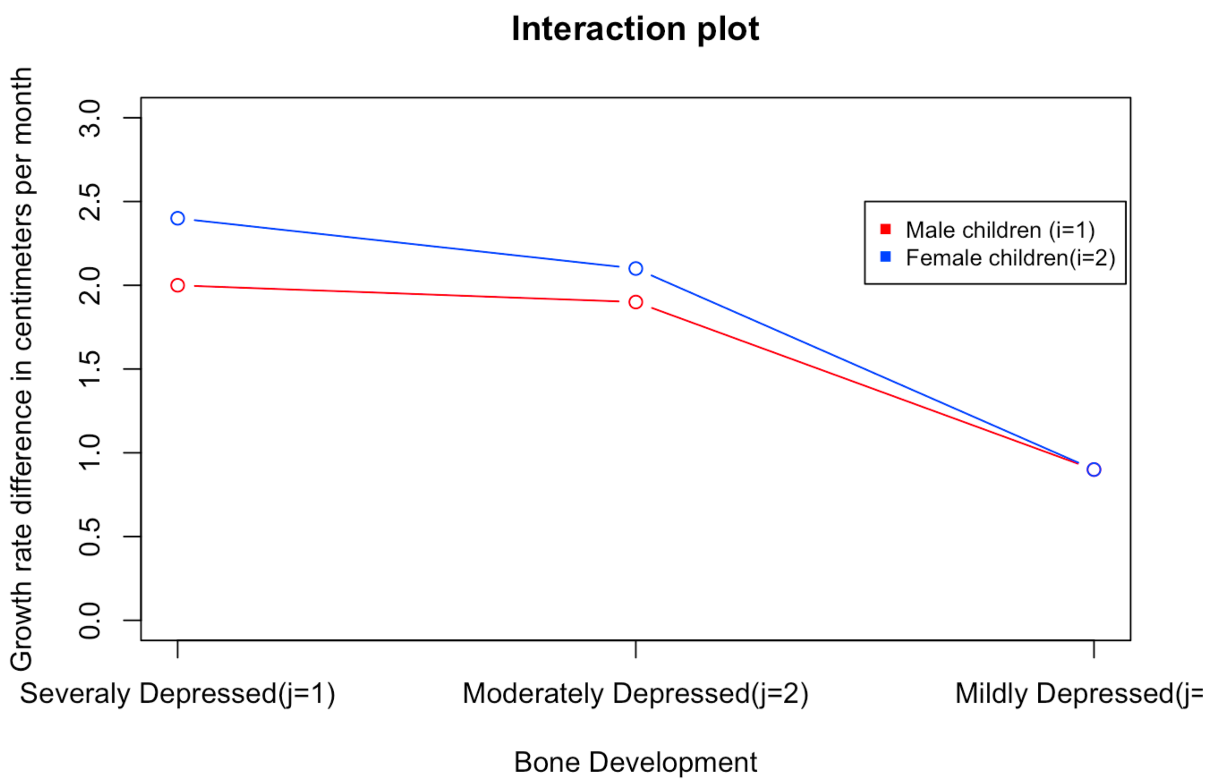
Decision rule:

If $F^* \leq F_{1-\alpha}(df_R - df_F, df_F)$, then conclude H_0

If $F^* > F_{1-\alpha}(df_R - df_F, df_F)$, then conclude H_a

Example

Plot the estimated treatment means $\bar{Y}_{ij}$ in the form of interaction plot. Does it appear that any factor effects are present? Explain.



Male children with severely depressed bone development benefit less during the growth hormone treatment than their female counterpart. This diffenrential effect tend to go away with mildly depressed bone development.It raises the quesitons whether some interaction effects are present.

It's clear that bone developement has a major impact on the change in growth rate during the growth hormone treatment. Severely depressed children seem to benefit more from it.

It's not clear whether gender of a child affects their reaction to the growth hormone treatment, as there is no clear sign of gender main effects.

Example

Test whether or not interaction effects are present by fitting the full and reduced regression models; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

Test for Interaction Effects: To test whether or not interaction effects are present

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = 0$$

H_a : not both $(\alpha\beta)_{11}$ and $(\alpha\beta)_{12}$ equal zero

we are simply testing whether or not two regression coefficients equal zero, using the generalized linear test approach.

$$Y_{ijk} = \mu_{..} + \alpha_1 X_{ijk1} + \beta_1 X_{ijk2} + \beta_2 X_{ijk3} + \varepsilon_{ijk} \qquad \text{Reduced model}$$

Code

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x1         1 0.0029  0.0029  0.0176 0.897785
## x2         1 3.6509  3.6509 22.4668 0.001464 **
## x3         1 0.7451  0.7451  4.5855 0.064638 .
## x1:x2      1 0.0754  0.0754  0.4642 0.514913
## x1:x3      1 0.0000  0.0000  0.0000 1.000000
## Residuals  8 1.3000  0.1625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 0.0029  0.0029  0.0208 0.8882630
## x2         1 3.6509  3.6509 26.5435 0.0004302 ***
## x3         1 0.7451  0.7451  5.4175 0.0422260 *
## Residuals 10 1.3754  0.1375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
## [1] 0.2320879
```

Code

```
## [1] 4.45897
```

To control the risk of making a Type 1 error at $\alpha = .05$, we require $F(.95; 2, 8) = 4.46$. Since $F^* = .23 \leq 4.46$, we conclude H_0, that no interaction effects are present.

Linear Regression Formulation to Two-Factor Studies with Unequal Sample Sizes

Tests for Factor Main Effects

To test whether factor A main effects are present:

$H_0 : \alpha_1 = \ldots = \alpha_a = 0 \qquad H_a : \text{not all } \alpha_i = 0$

Full model
$$Y_{ijk} = \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \ldots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk;.,1} + \ldots + \beta_{b-1} X_{ijk;.,b-1}}_{\text{B main effect}}$$

Reduced model
$$Y_{ijk} = \mu_{..} + \underbrace{\beta_1 X_{ijk;.,1} + \ldots + \beta_{b-1} X_{ijk;.,b-1}}_{\text{B main effect}}$$

$$F^* = \frac{\text{SSE}(R) - \text{SSE}(F)}{df_R - df_F} \div \frac{\text{SSE}(F)}{df_F}$$

Decision rule:

If $F^* \leq F_{1-\alpha}(df_R - df_F, df_F)$, then conclude H_0

If $F^* > F_{1-\alpha}(df_R - df_F, df_F)$, then conclude H_a

To test whether factor B main effects are present:

$H_0 : \beta_1 = \ldots = \beta_b = 0 \qquad H_a : \text{not all } \beta_j = 0$

Full model
$$Y_{ijk} = \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \ldots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk;.,1} + \ldots + \beta_{b-1} X_{ijk;.,b-1}}_{\text{B main effect}}$$

Reduced model
$$Y_{ijk} = \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \ldots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}}$$

$$F^* = \frac{\text{SSE}(R) - \text{SSE}(F)}{df_R - df_F} \div \frac{\text{SSE}(F)}{df_F}$$

Decision rule:

If $F^* \leq F_{1-\alpha}(df_R - df_F, df_F)$, then conclude H_0

If $F^* > F_{1-\alpha}(df_R - df_F, df_F)$, then conclude H_a

Example

State the reduced regression models for testing for subject matter and highest degree main effects, respectively, and conduct each of the tests. Use $\alpha = .05$ each time and state the alternatives, decision rule, and conclusion.

Test for Factor A Main Effect.

$$H_0 : \alpha_1 = 0$$
$$H_a : \alpha_1 \neq 0$$

$$Y_{ijk} = \mu_{...} + \beta_1 X_{ijk2} + \beta_2 X_{ijk3} + (\alpha\beta)_{11} X_{ijk!} X_{ijk2} + (\alpha\beta)_{12} X_{ijk1} X_{ijk3} + \varepsilon_{ijk} \quad \text{Reduced model}$$

Code

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x1         1 0.0029  0.0029  0.0176 0.897785
## x2         1 3.6509  3.6509 22.4668 0.001464 **
## x3         1 0.7451  0.7451  4.5855 0.064638 .
## x1:x2      1 0.0754  0.0754  0.4642 0.514913
## x1:x3      1 0.0000  0.0000  0.0000 1.000000
## Residuals  8 1.3000  0.1625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x2         1 3.4410  3.4410 21.8092 0.001169 **
## x3         1 0.8653  0.8653  5.4842 0.043889 *
## x4         1 0.0462  0.0462  0.2925 0.601735
## x5         1 0.0018  0.0018  0.0117 0.916233
## Residuals  9 1.4200  0.1578
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
## [1] 0.7384615
```

Code

```
## [1] 4.45897
```

Example

Test for Factor B Main Effect. $H_0: \beta_1 = \beta_2 = 0$ H_a : not both β_j equal zero

$$\begin{aligned} Y_{ijk} =& \mu_{..} + \alpha_1 X_{ijk|} + (\alpha\beta)_{11} X_{ijk|} X_{ijk2} \\ &+ (\alpha\beta)_{12} X_{ijk|} X_{ijk.3} + \varepsilon_{ijk} \qquad \text{Reduced model} \end{aligned}$$

Code

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x1         1 0.0029  0.0029  0.0176 0.897785
## x2         1 3.6509  3.6509 22.4668 0.001464 **
## x3         1 0.7451  0.7451  4.5855 0.064638 .
## x1:x2      1 0.0754  0.0754  0.4642 0.514913
## x1:x3      1 0.0000  0.0000  0.0000 1.000000
## Residuals  8 1.3000  0.1625
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## x1         1 0.0029 0.00286  0.0052 0.9439
## x4         1 0.0207 0.02071  0.0377 0.8499
## x5         1 0.2610 0.26100  0.4754 0.5062
## Residuals 10 5.4897 0.54897
```

Code

```
## [1] 12.89143
```

Code

```
## [1] 4.45897
```

we conclude that there are no factor A main effects but that factor B main effects are present.

Thus, these tests support the indications obtained previously from the estimated treatment means plot, that a child's bone development affects the change in growth rate during growth hormone treatment and that there are no gender and interaction effects.

Example

Make all pairwise comparisons between the bone development; use the Tukey procedure with a 90 percent family confidence coefficient. State your findings.

Since $\mu_{.j.} = \frac{\sum_i \mu_{ij}}{a}$

$$\hat{\mu}_{.j} = \frac{\sum_i \bar{Y}_{ij}}{a}$$

$$s^2\{\hat{\mu}_{.j}\cdot\} = \frac{MSE}{a^2}\sum_i \frac{1}{n_{ij}}$$

The pairwise comparison for differnet bone development groups :

$$\hat{D}_1 = \hat{\mu}_{.1} - \hat{\mu}_{.2} : \hat{\mu}_{.1} - \hat{\mu}_{.2} \pm \frac{1}{\sqrt{2}} q(.90; b, \frac{\sum n_{ij}}{ab} ab) \frac{MSE}{a^2} \sum_i (\frac{1}{n_{i1}} + \frac{1}{n_{i2}})$$

Code

```
## [1] -0.5078135  0.9078135
```

$$\hat{D}_2 = \hat{\mu}_{.1} - \hat{\mu}_{.3} : \hat{\mu}_{.1} - \hat{\mu}_{.3} \pm \frac{1}{\sqrt{2}} q(.90; b, \frac{\sum n_{ij}}{ab} ab) \frac{MSE}{a^2} \sum_i (\frac{1}{n_{i1}} + \frac{1}{n_{i3}})$$

Code

```
## [1] 0.5921865 2.0078135
```

$$\hat{D}_3 = \hat{\mu}_{.2} - \hat{\mu}_{.3} : \hat{\mu}_{.2} - \hat{\mu}_{.3} \pm \frac{1}{\sqrt{2}} q(.90; b, \frac{\sum n_{ij}}{ab} ab) \frac{MSE}{a^2} \sum_i (\frac{1}{n_{i2}} + \frac{1}{n_{i3}})$$

Code

```
## [1] 0.4792065 1.7207935
```

We conclude from these confidence intervals with 90 percent family confidence coefficient that among children with growth deficiency:

- children with only mildly depressed bone development (less severe growth deficiency) on the average have a substantially smaller increase in the growth rate than children with either moderately depressed or severely depressed bone development.
- Further, the latter two groups of children do not show significantly different mean changes in the growth rate.