

# **Lecture 6:**

# **Linear Regression Approach to**

# **ANOVA**

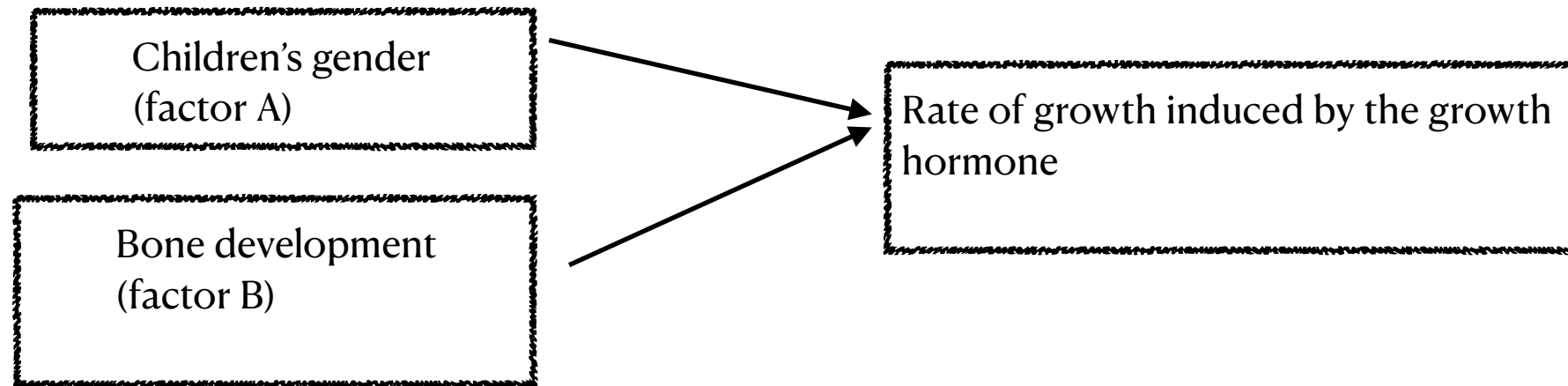
**STA 106: Analysis of Variance**

# Example

(The Growth Hormone Study)

The objective of the study:

Synthetic growth hormone was administered at a clinical research center for those children with growth hormone deficiency.



The study setup:

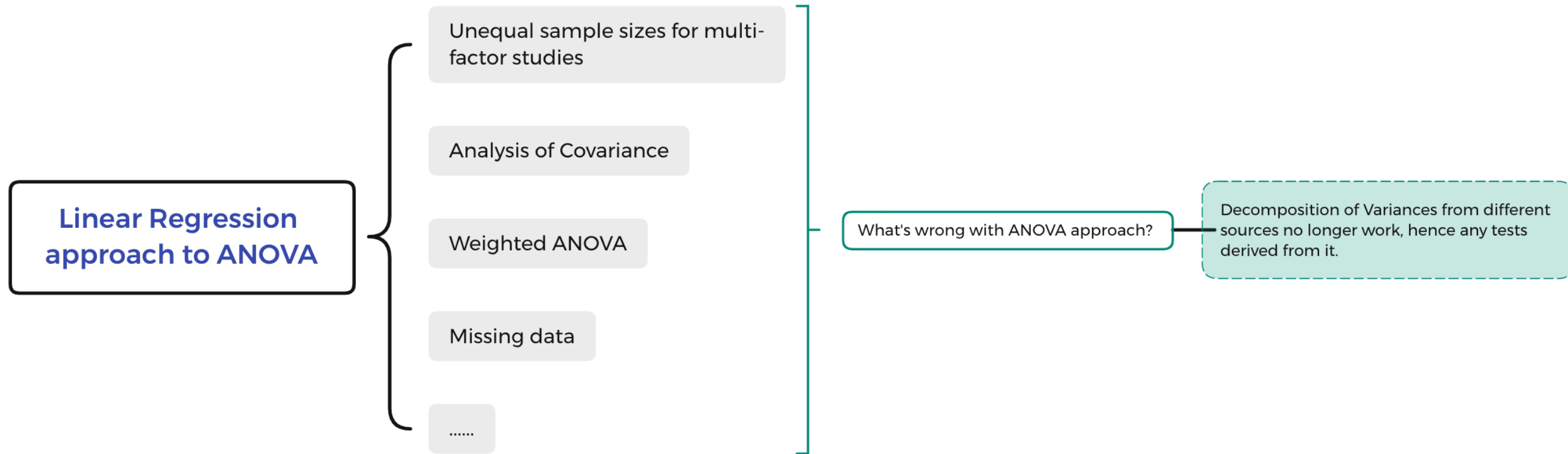
3 children were randomly selected for each bone-gender combination

Y: growth rate during hormone treatment - growth rate prior

but 4 children were unable to complete the study —> unequal sample size across treatments

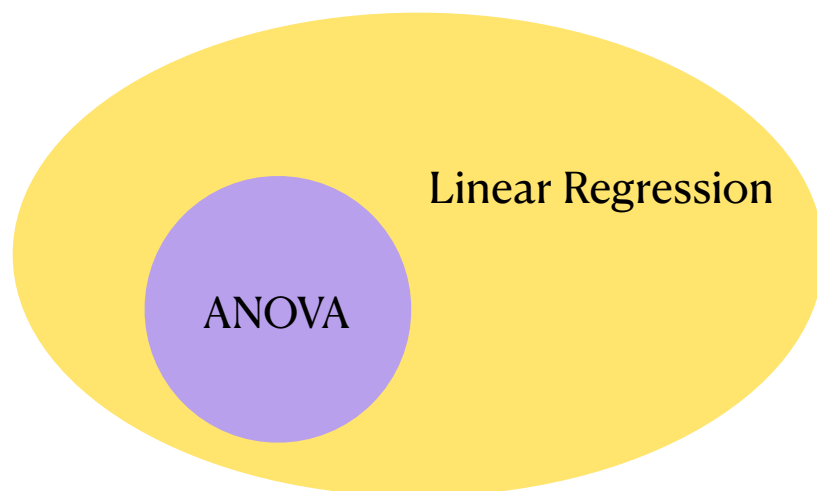
Gender (factor A) <i>i</i>	Bone Development (factor B) <i>j</i>		
	Severely Depressed ( <i>B</i> <sub>1</sub> )	Moderately Depressed ( <i>B</i> <sub>2</sub> )	Mildly Depressed ( <i>B</i> <sub>3</sub> )
Male ( <i>A</i> <sub>1</sub> )	1.4 ( <i>Y</i> <sub>111</sub> )	2.1 ( <i>Y</i> <sub>121</sub> )	.7 ( <i>Y</i> <sub>131</sub> )
	2.4 ( <i>Y</i> <sub>112</sub> )	1.7 ( <i>Y</i> <sub>122</sub> )	1.1 ( <i>Y</i> <sub>132</sub> )
	2.2 ( <i>Y</i> <sub>113</sub> )		
	Mean 2.0 ( $\bar{Y}_{11\cdot}$ )	1.9 ( $\bar{Y}_{12\cdot}$ )	.9 ( $\bar{Y}_{13\cdot}$ )
Female ( <i>A</i> <sub>2</sub> )	2.4 ( <i>Y</i> <sub>211</sub> )	2.5 ( <i>Y</i> <sub>221</sub> )	.5 ( <i>Y</i> <sub>231</sub> )
		1.8 ( <i>Y</i> <sub>222</sub> )	.9 ( <i>Y</i> <sub>232</sub> )
		2.0 ( <i>Y</i> <sub>223</sub> )	1.3 ( <i>Y</i> <sub>233</sub> )
	Mean 2.4 ( $\bar{Y}_{21\cdot}$ )	2.1 ( $\bar{Y}_{22\cdot}$ )	.9 ( $\bar{Y}_{23\cdot}$ )

# Why Linear Regression Approach to ANOVA?



In fact, ANOVA Model is a specialized **linear model** for experimental data ( originally ).

**Linear model = Linear Regression model**



# Linear Regression Approach to ANOVA



Regression Formulation to Single-Factor Studies

Linear Regression Formulation to Two-Factor Studies

Analysis of Covariance Model (ANCOVA)

# Regression Formulation to Single-Factor Studies

For single-factor studies, there is no difference between ANOVA and Linear Regression approach

There is no need to use linear regression since single-factor ANOVA is extremely simple, due to simple X matrix structure.

How ANOVA models can be written as a linear regression model?

One-Way ANOVA Model  $Y_{ij} = \mu_i + \varepsilon_{ij}$

Define  $i$ th treatment effect:  $\tau_i = \mu_i - \mu_{\cdot}$

Define unweighted average of all treatment means:  $\mu_{\cdot} = \frac{\sum_{i=1}^r \mu_i}{r}$

constraint:  $\sum_{i=1}^r \tau_i = 0$

$\Rightarrow \tau_r = -\tau_1 - \dots - \tau_{r-1}$

$\Rightarrow Y_{ij} = \mu_{\cdot} + \tau_i + \varepsilon_{ij}$

# Regression Formulation to Single-Factor Studies



$$Y_{ij} = \mu. + \tau_i + \varepsilon_{ij}$$

Define indicator variables used in linear regression:

$$X_{ij,1} = \begin{cases} 1 & \text{if case from level 1} \\ -1 & \text{if case from level } r \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij,2} = \begin{cases} 1 & \text{if case from level 2} \\ -1 & \text{if case from level } r \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij,r-1} = \begin{cases} 1 & \text{if case from level } r-1 \\ -1 & \text{if case from level } r \\ 0 & \text{otherwise} \end{cases}$$



One-Way ANOVA Model can be written as Linear Regression Model:

$$Y_{ij} = \mu. + \tau_1 X_{ij,1} + \dots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

Diagram illustrating the components of the equation:

- $Y_{ij}$  is labeled as the **Dependent variable**.
- $X_{ij,1}$  and  $X_{ij,r-1}$  are labeled as **independent variable**.

Response is a linear combination of parameters  $\mu., \tau_1, \dots, \tau_{r-1}$

# Regression Formulation to Single-Factor Studies



One-Way ANOVA Model can be written as Linear Regression Model:

$$Y_{ij} = \mu_{\cdot} + \tau_1 X_{ij,1} + \dots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

$$\text{Regression parameters} \quad \begin{cases} \bar{\mu}_{\cdot} & : \text{intercept} \\ \tau_1 \dots \tau_{r-1} & : \text{slope coefficients} \end{cases}$$

Example:  $r = 3, n_1 = n_2 = n_3 = 2$

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_{\cdot} \\ \tau_1 \\ \tau_2 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{21} \\ \varepsilon_{22} \\ \varepsilon_{31} \\ \varepsilon_{32} \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

# Regression Formulation to Single-Factor Studies

General Linear Test Approach: test about regression parameters

Test for equality of factor level means = all treatment effects are zero:

$$H_0 : \tau_1 = \dots = \tau_{r-1} = 0$$



restricted model

$$Y_{ij} = \mu_{.} + \varepsilon_{ij}$$

$$SSE(R) = SSTO$$

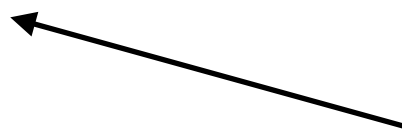
$$H_a : \text{not all } \tau_i\text{'s equal to zero}$$



Full or unrestricted model

$$Y_{ij} = \mu_{.} + \tau_1 X_{ij,1} + \dots + \tau_{r-1} X_{ij,r-1} + \varepsilon_{ij}$$

$$SSE(F) = SSE$$



Idea:

More parameters included in the model  $\rightarrow$  better one can fit the data  $\rightarrow$  smaller error variance

To compare the two SSE's



# Regression Formulation to Single-Factor Studies

General Linear Test Approach: test about regression parameters

$SSE(F) \approx SSE(R)$ :


using extra parameters in the full model does not account for much more variability than the reduced model, in which case the data suggest that reduced model is equally adequate, so the extra parameters should in fact be negligible.

This favors  $H_0$ , small difference  $SSE(R) - SSE(F)$  favors  $H_0$

$SSE(F) << SSE(R)$ :

Extra parameters do help to reduce variation, therefore should be included in the model.

This favors  $H_a$ , large difference  $SSE(R) - SSE(F)$  favors  $H_a$


$$F^* = \frac{\frac{SSE(R) - SSE(F)}{df_R - df_F}}{\frac{SSE(F)}{df_F}} \sim F(df_R - df_F, df_F)$$
$$= \frac{\frac{SSTR}{df_{SSTR}}}{\frac{SSE}{df_{SSE}}} = \frac{MSTR}{MSE} \quad \text{ANOVA's test for equality of factor level means!}$$

# 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

by design

illness of subjects

lost to follow up

technical problems

.....

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

## Analysis of Variance

Partition of Total Sum of Squares

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

Total deviation

Deviation of estimated  
treatment mean around  
overall mean

Deviation around  
estimated treatment mean

$$\sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{...})^2 = \sum_i \sum_j \sum_k (\bar{Y}_{ij.} - \bar{Y}_{...})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2$$

Total variation

$$= n \sum_i \sum_j (\bar{Y}_{ij.} - \bar{Y}_{...})^2 + \sum_i \sum_j \sum_k (Y_{ijk} - \bar{Y}_{ij.})^2$$

Variation due to factor A and B

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

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

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



→ SSTO = SSTR + SSE

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

## Analysis of Variance

Partition of Treatment Sum of Squares.

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

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

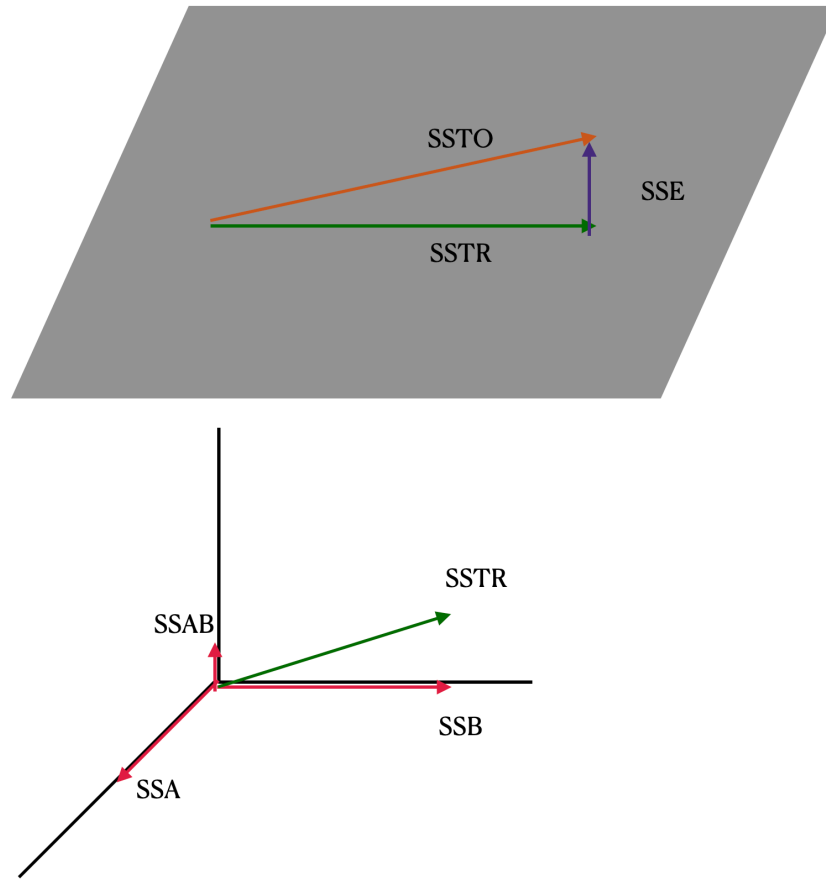
➡

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

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

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

$$\bullet \mu_{..} = \frac{\sum_{i=1}^a \sum_{j=1}^b \mu_{ij}}{ab} : \text{overall mean}$$

$$\bullet \alpha_i = \mu_{i.} - \mu_{..} \text{ main effect of factor A at } i\text{th level}$$

$$\text{Subject to (a-1) constraints } \sum \alpha_i = 0$$

$$\bullet \beta_j = \mu_{.j} - \mu_{..} \text{ main effect of factor B at } j\text{th level}$$

$$\text{Subject to (b-1) constraints } \sum \beta_j = 0$$

$$\bullet \gamma_{ij} = \mu_{ij} - \alpha_i - \beta_j = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..} \text{ interaction effect of factor A at } i\text{th level with factor B at } j\text{th level}$$

$$\text{Subject to a+b-1 constraints}$$

$$\sum_i (\alpha\beta)_{ij} = 0 \quad j = 1, \dots, b$$

$$\sum_j (\alpha\beta)_{ij} = 0 \quad i = 1, \dots, a$$

$$\bullet \varepsilon_{ijk} \text{ are independent } N(0, \sigma^2) \text{ for } i = 1, \dots, a; j = 1, \dots, b; k = 1, \dots, n_{ij}$$

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

Indicator variables for factor A

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

...

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

---

Indicator variables for factor B

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

...

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

➔

$$Y_{ijk} = \mu_{..} + \underbrace{\alpha_1 X_{ijk;1,.} + \dots + \alpha_{a-1} X_{ijk;a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk,.,1} + \dots + \beta_{b-1} X_{ijk,.,b-1}}_{\text{B main effect}} + \underbrace{\gamma_{11} X_{ijk;1,.} X_{ijk,.,1} + \dots + \gamma_{(a-1)(b-1)} X_{ijk;a-1,.} X_{ijk,.,b-1}}_{\text{AB interaction effect}} + \varepsilon_{ijk}$$

Regression coefficients  $\mu_{..}, \alpha_1, \dots, \alpha_{a-1}, \beta_1, \dots, \beta_{b-1}, \gamma_{11}, \dots, \gamma_{(a-1)(b-1)}$  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

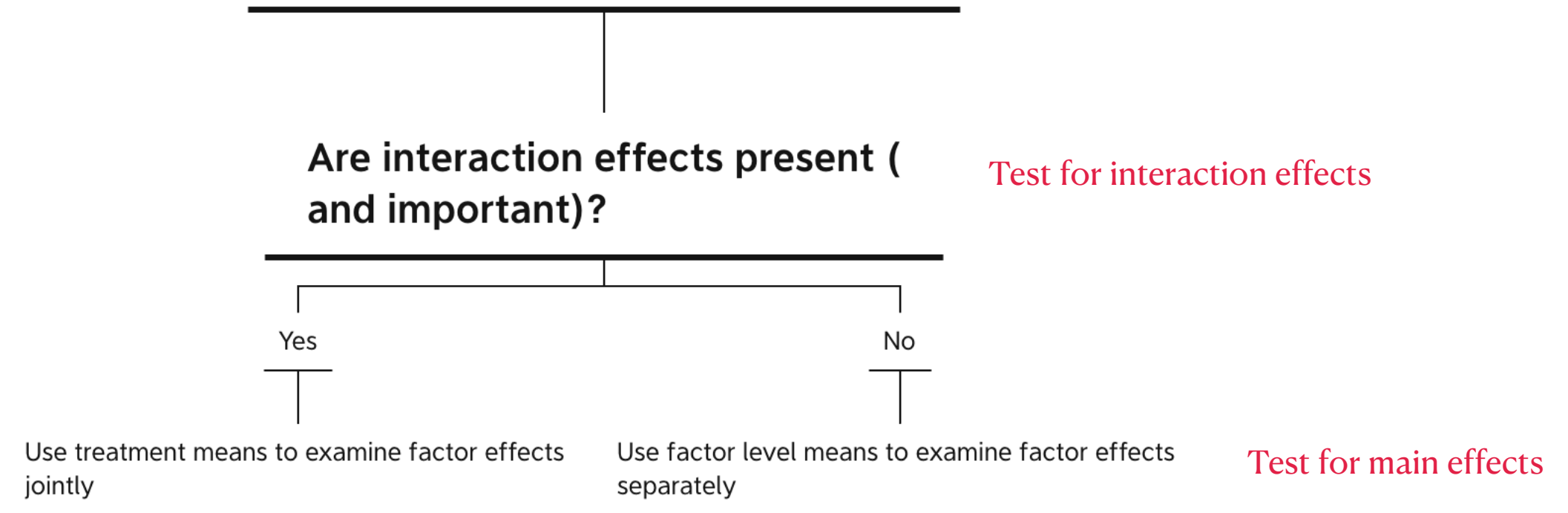
So we need more general analysis of variance approach in the regression analysis :

## **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

## Strategy for Analysis of Two-Factor Studies



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

## Test for Interaction Effects

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

$$\begin{aligned} \text{Full model} \quad Y_{ijk} = & \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \dots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk,.,1} + \dots + \beta_{b-1} X_{ijk,.,b-1}}_{\text{B main effect}} \\ & + \underbrace{\gamma_{11} X_{ijk,1,.} X_{ijk,.,1} + \dots + \gamma_{(a-1)(b-1)} X_{ijk,a-1,.} X_{ijk,.,b-1}}_{\text{AB interaction effect}} + \varepsilon_{ijk} \end{aligned}$$

$$\begin{aligned} \text{Reduced model} \quad Y_{ijk} = & \mu_{..} + \underbrace{\alpha_1 X_{ijk,1,.} + \dots + \alpha_{a-1} X_{ijk,a-1,.}}_{\text{A main effect}} + \underbrace{\beta_1 X_{ijk,.,1} + \dots + \beta_{b-1} X_{ijk,.,b-1}}_{\text{B main effect}} \end{aligned}$$

F test:

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

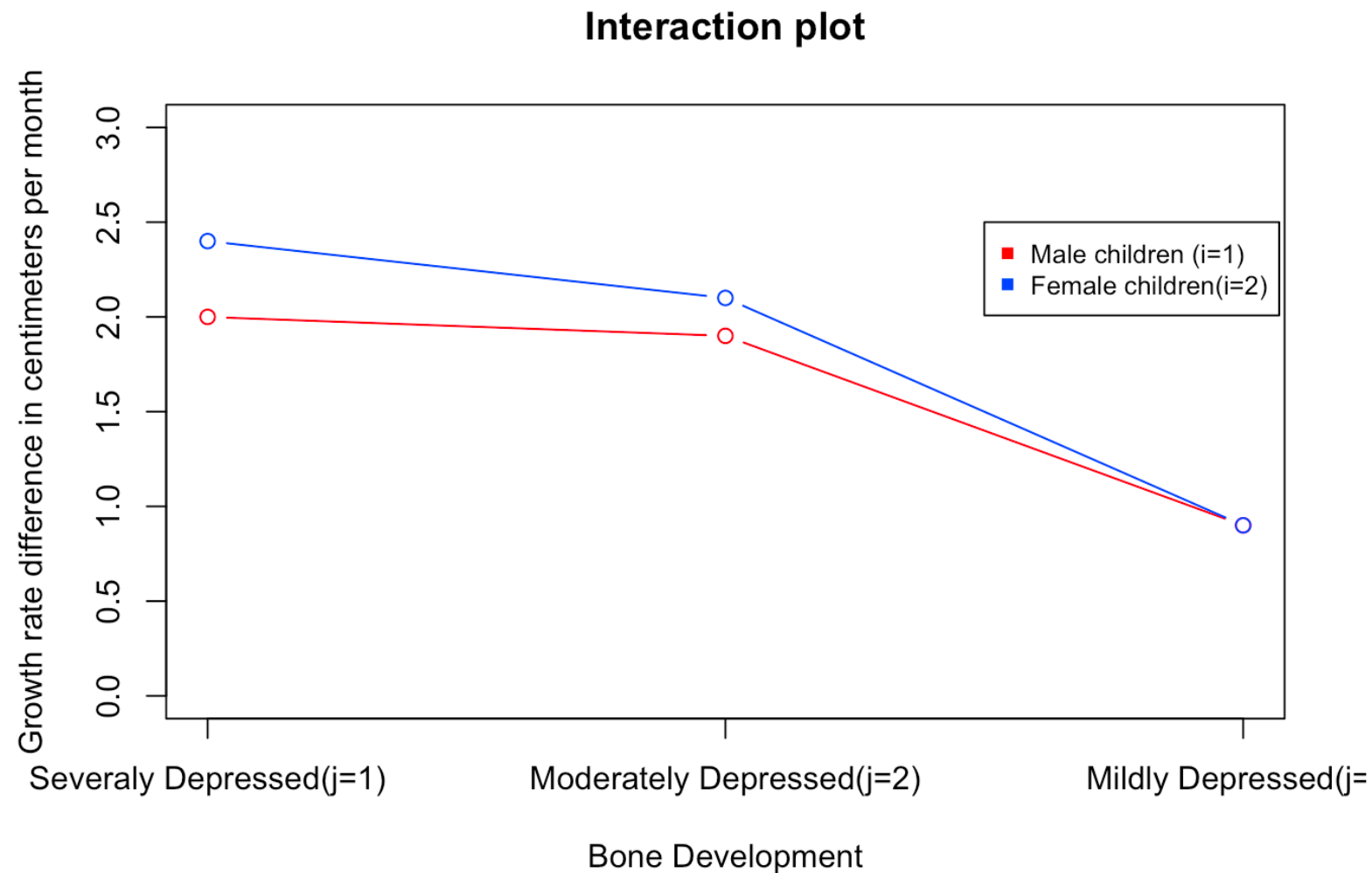
Decision rule:

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 differential effect tends to go away with mildly depressed bone development. It raises the questions whether some interaction effects are present.

It's clear that bone development 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.

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

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)
## 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:

$$H_0 : \alpha_1 = \dots = \alpha_a = 0 \quad H_a : \text{not all } \alpha_i = 0$$

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

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

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

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:

$$H_0 : \beta_1 = \dots = \beta_b = 0 \quad H_a : \text{not all } \beta_j = 0$$

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

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

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

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.

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

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

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  $\beta_j$  equal zero

$$Y_{ijk} = \mu_{..} + \alpha_1 X_{ijk1} + (\alpha\beta)_{11} X_{ijk1} X_{ijk2} \\ + (\alpha\beta)_{12} X_{ijk1} X_{ijk3} + \varepsilon_{ijk} \quad \text{Reduced model}$$

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

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

The pairwise comparison for different bone development groups :

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

Code

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

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

Code

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

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

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.

# Linear Regression Approach to ANOVA

Regression Formulation to Single-Factor Studies

Linear Regression Formulation to Two-Factor Studies

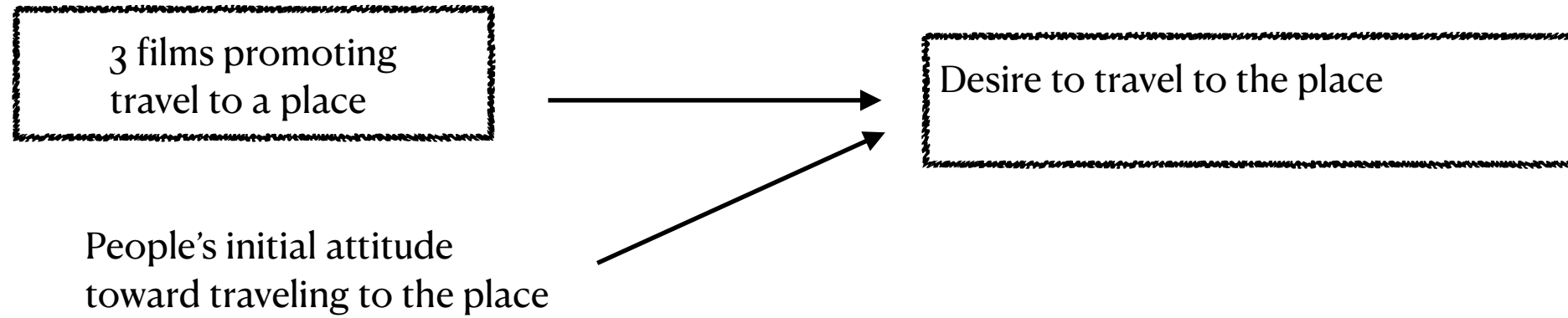


Analysis of Covariance Model (ANCOVA)

# Example

(The Travel Promotion Study)

The objective of the study:



The study setup:

15 participants were randomly and equally split into 3 groups, with each group viewing one film.

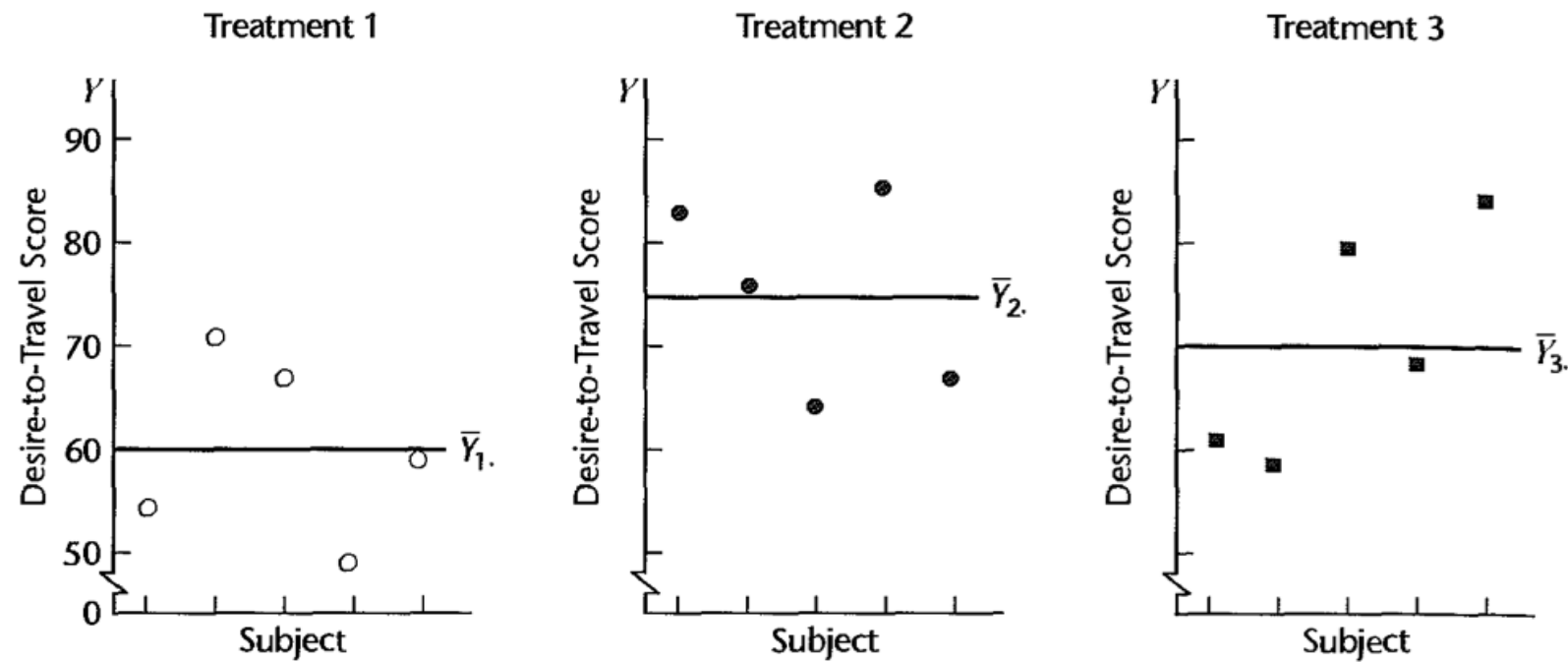
Their initial attitude toward traveling to the place and the desire after viewing the assigned promotion film were recorded.

☒ What type of study?

# Example

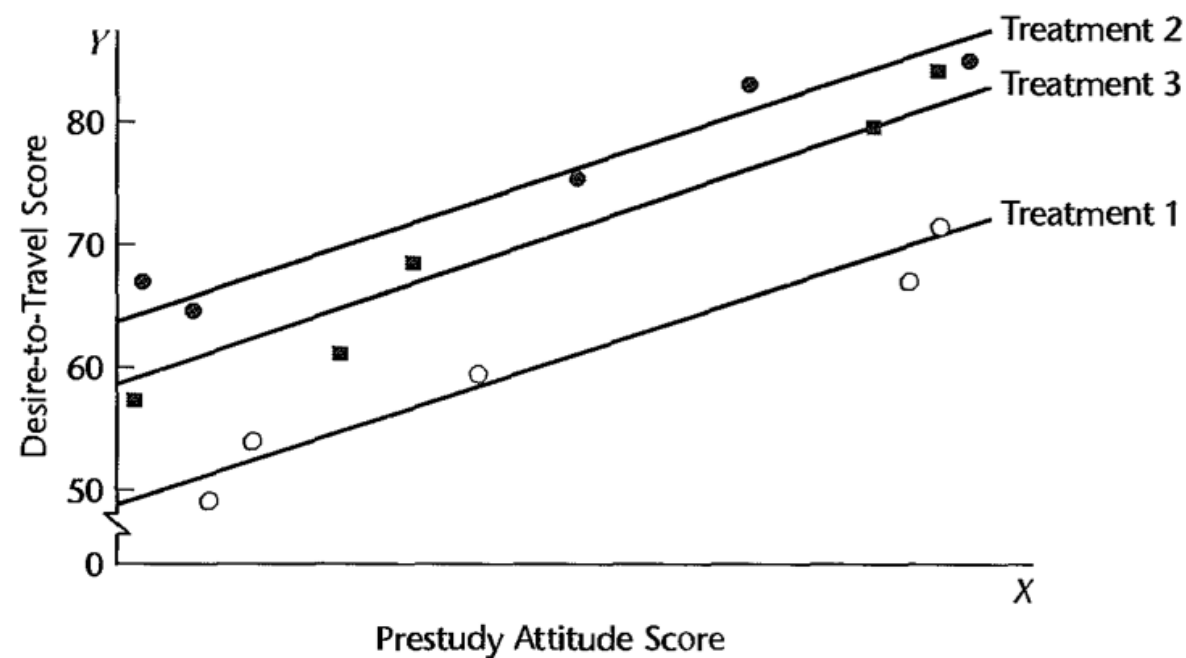
(The Travel Promotion Study)

(a) Error Variability with Single-factor Analysis of Variance Model



Error term: scatter around the treatment means are fairly large

(b) Error Variability with Covariance Analysis Model



Prestudy attitude is highly suggestive of post-study attitude, in fact, they are highly linearly associated.

If we incorporate pre study attitude, the error variability is much smaller, which will lead to more precise inference.

# Analysis of Covariance (ANCOVA)

## ANCOVA

Utilizes the relationship between the response variables and one or more variables, to  
reduce the error term variability,  
make more precision inference,  
and make the study a more powerful one for comparing treatment effects

## Concomitant variables

We call variables added to ANOVA model in order to reduce error term variability:  
**Concomitant variable or Covariate** (natural accompanying or occurring)

Must have relation to the response, otherwise, no good

Human subjects: pre-study score, age, socioeconomic status, aptitude....

## Difference between Blocking versus ANCOVA?

Blocking is used in the design stage, when we have one or more qualitative variables, to reduce the error term variability

ANCOVA is used in the analysis stage

# Single-Factor ANCOVA Model

ANCOVA model starts with ANOVA model, adds one more term reflecting the relation between Y and concomitant variable X:

One-Way ANOVA Model  $Y_{ij} = \mu. + \tau_i + \epsilon_{ij}$



$$Y_{ij} = \mu. + \tau_i + \beta X_{ij} + \epsilon_{ij}$$



Regression coefficient describing relation  
between Y ~ X

Centered version:  $Y_{ij} = \mu. + \tau_i + \beta(X_{ij} - \bar{X}_{..}) + \epsilon_{ij}$

# Single-Factor ANCOVA Model

ANCOVA Model: 
$$Y_{ij} = \mu. + \tau_1 I_{ij,1} + \dots + \tau_{r-1} I_{ij,r-1} + \beta X_{ij} + \varepsilon_{ij}$$

Where 
$$I_1 = \begin{cases} 1 & \text{if store received treatment 1} \\ -1 & \text{if store received treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

$$I_{r-1} = \begin{cases} 1 & \text{if store received treatment } r - 1 \\ -1 & \text{if store received treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij} = X_{ij} - \bar{X}_{..}$$

There is no interest in whether the regression coefficient beta is zero or not, we are not interested in finding out its relation with Y.

We just want to use this relation to reduce error variability.

If beta is indeed not zero, i.e. there is a regression relation between X and Y, good

If not, no bias is introduced

# Single-Factor ANCOVA Model

## Comparisons of Treatment Effects

$$\text{ANOVA: } E(\eta_{ij}) = \mu_i = \mu. + \tau_i$$

All observations for ith treatment has the same mean response, therefore comparing treatment effect is just comparing treatment means

$$\text{ANCOVA: } E(\eta_{ij}) = \mu. + \tau_i + \beta(X_{ij} - \bar{X}_{..})$$

Mean response not only depend on the treatment, but also on concomitant variable X



A meaningful measure the effect of treatment 1 versus treatment 2 have to let the value of concomitant X to be the same:

Treatment effect comparison:  $\tau_1 - \tau_2$

Measures how much higher the mean response is with treatment 1 versus 2 for any fixed value of X

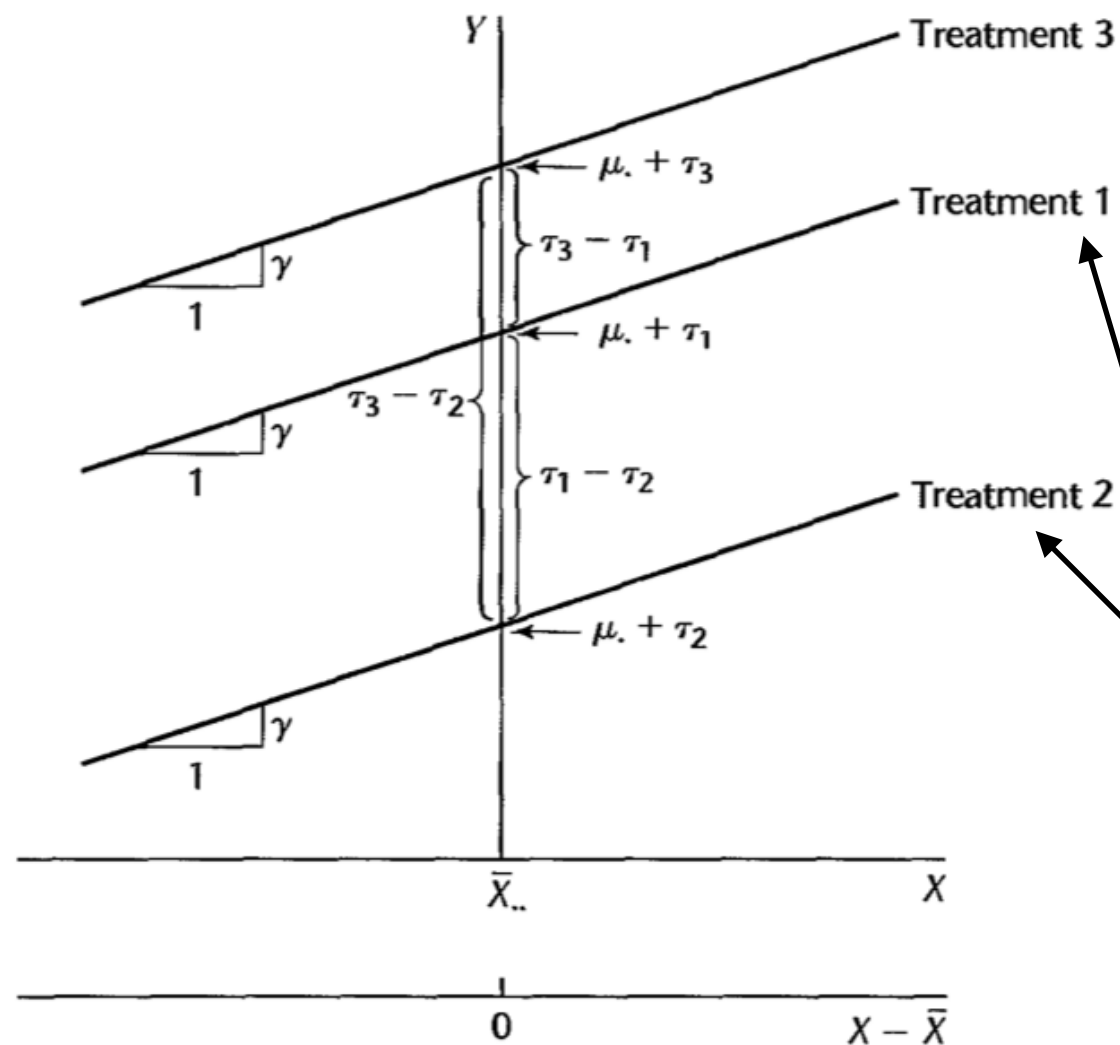


To test for the presence of treatment or factor effect:

$$H_0 : \tau_1 = \dots = \tau_r = 0 \text{ vs } H_a : \text{ not all } \tau_i \text{ equal zero}$$



# Single-Factor ANCOVA Model



We can no longer speak of the mean response for  $i$ th treatment, as it varies with  $X$

## Compare “treatment effect”:

Treatment 1 leads to a higher mean response than treatment 2 by a fixed amount, regardless of what the value  $X$  is.

The constant differential effect is reflected by the parallel lines.

If there is no differential treatments, then all lines must be identical. Otherwise, the treatment effects are reflected by the vertical distances.

# Example

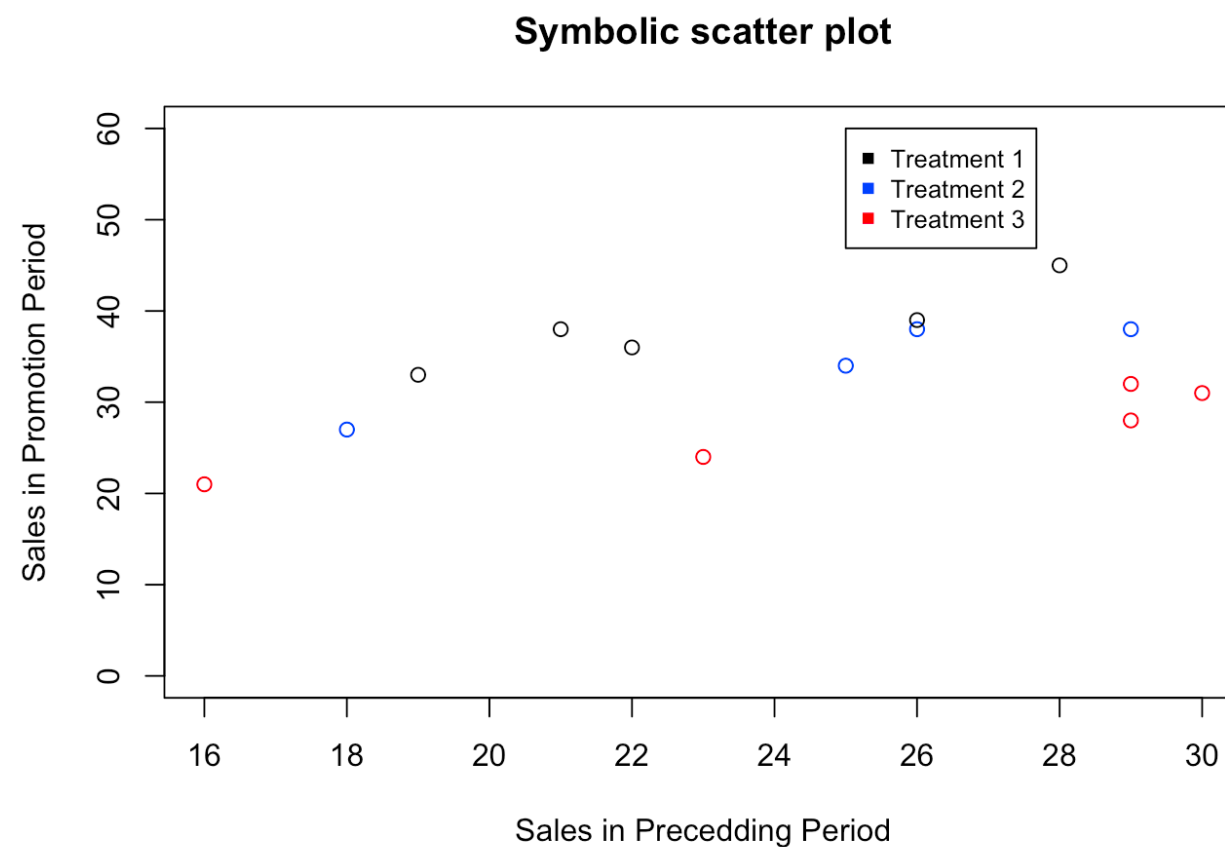
A company studied the effects of three different types of promotions on sales of its crackers:

- Treatment 1-Simpling of product by customers in store and regular shelf space
- Treatment 2-Additional shelf space in regular location
- Treatment 3-Special display shelves at ends of aisle in addition to regular shelf space

Fifteen stores were selected for the study, and a completely randomized experimental design was utilized. Each store was randomly assigned one of the promotion types, with five stores assigned to each type of promotion. Other relevant conditions under the control of the company, such as price and advertising, were kept the same for all stores in the study. Data on the number of cases of the product sold during the promotional period, denoted by  $Y$ , as are also data on the sales of the product in the preceding period, denoted by  $X$ . Sales in the preceding period are to be used as the concomitant variable. Assume that covariance model is applicable.

Does it appear that sales in preceding period is a good concomitant variable?

Does it appear that there are treatment effects?



Sales in preceding period seems to have strong linear relationship with sales in promotion period, therefore it is a good concomitant variable.

Treatment 1 seems to have highest average sales in promotion period, followed by treatment 2, then treatment 1, there appears to be some treatment effects.

## Example

State the regression model equivalent to covariance model for this case; use 1,-1,0 indicator variables. Also state the reduced regression model for testing for treatment effects.

ANCOVA Model: 
$$Y_{ij} = \mu. + \tau_1 I_{ij1} + \tau_2 I_{ij2} + \gamma X_{ij} + \varepsilon_{ij}$$

Where 
$$I_1 = \begin{cases} 1 & \text{if store received treatment 1} \\ -1 & \text{if store received treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{if store received treatment 2} \\ -1 & \text{if store received treatment 3} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{ij} = X_{ij} - \bar{X}_{..}$$

# Example

Fit the full and reduced regression models and test for treatment effects; use  $\alpha = .01$ . State the alternatives, decision rule, and conclusion.

$H_0 : \tau_1 = \tau_2 = 0$  vs  $H_a : \text{not both } \tau_1 \text{ and } \tau_2 \text{ equal zero}$

$Y_{ij} = \mu. + \tau_1 I_{ij1} + \tau_2 I_{ij2} + \gamma X_{ij} + \varepsilon_{ij}$  Full model

$Y_{ij} = \mu. + \gamma X_{ij} + \varepsilon_{ij}$  Reduced model

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

```
## Analysis of Variance Table
##
## Response: y
##          Df Sum Sq Mean Sq F value    Pr(>F)
## I1          1 302.500  302.500   86.269 1.538e-06 ***
## I2          1  36.300   36.300   10.352 0.008196 **
## x           1 269.029  269.029   76.723 2.731e-06 ***
## Residuals 11  38.571    3.506
## ---
## 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)
## x           1 190.68  190.678    5.4393 0.03641 *
## Residuals 13 455.72   35.056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[Code](#)

```
## [1] 59.48282
```

[Code](#)

```
## [1] 3.982298
```

[Code](#)

The level of significance is to be controlled at  $\alpha = .05$ ; hence, we need to obtain  $F(.95; 2, 11) = 3.98$ . The decision rule therefore is: If  $F^* \leq 3.98$ , conclude  $H_0$  If  $F^* > 3.98$ , conclude  $H_a$  Since  $F^* = 59.5 > 3.98$ , we conclude  $H_a$ , that the three cracker promotions differ in sales effectiveness.

