Lecture 6: Linear Regression Approach to ANOVA

STA 106: Analysis of Variance

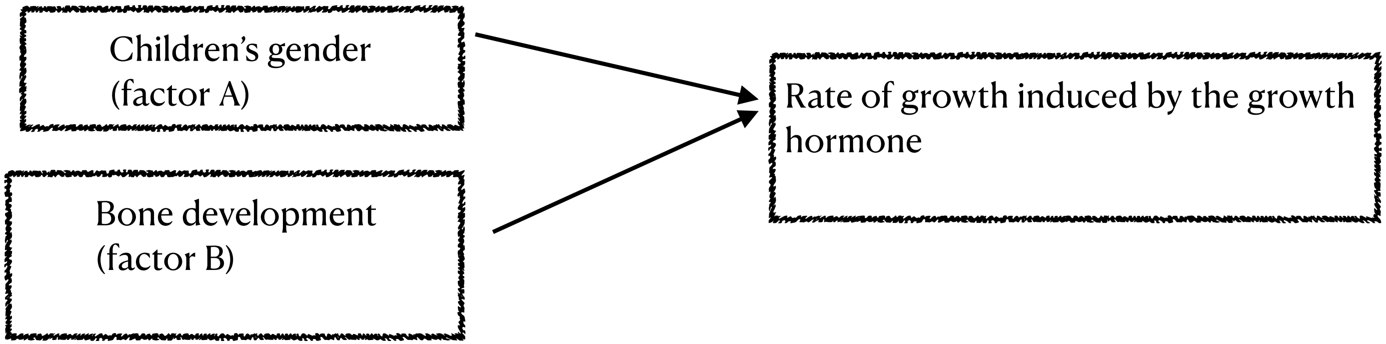
Xiner Zhou
Department of Statistics, University of California, Davis

Example

(The Growth Hormone Study)

<u>The objective of the study:</u>

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



<u>The study setup:</u>

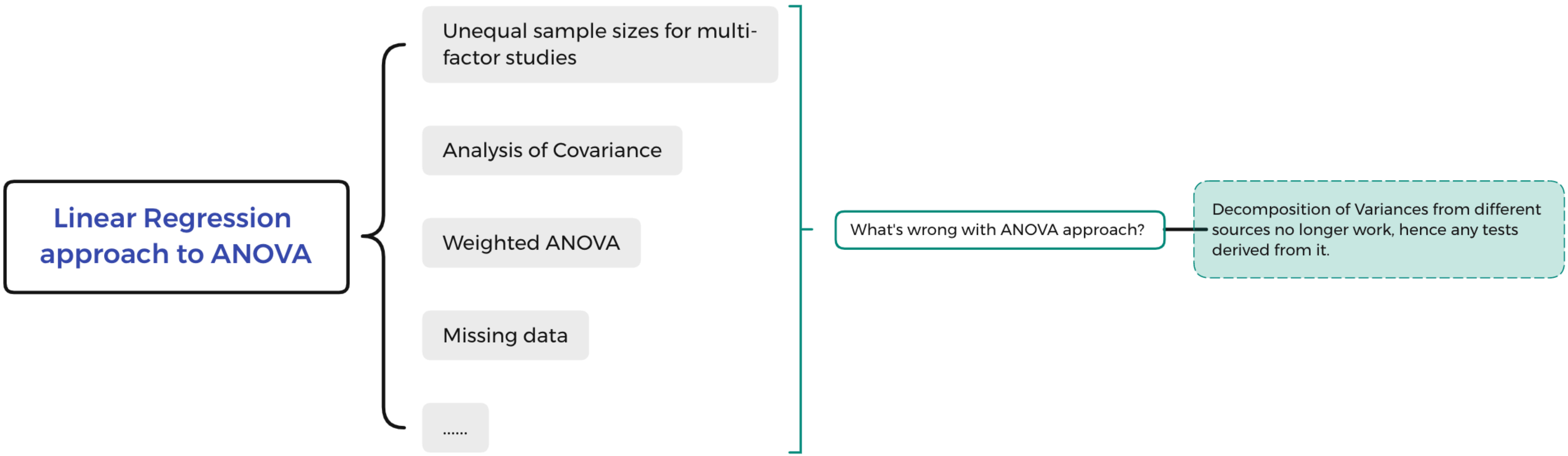
3 children were randomly selected for each bone-gender combination

Y: growth rate during hormone treatment - growth rate prior

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

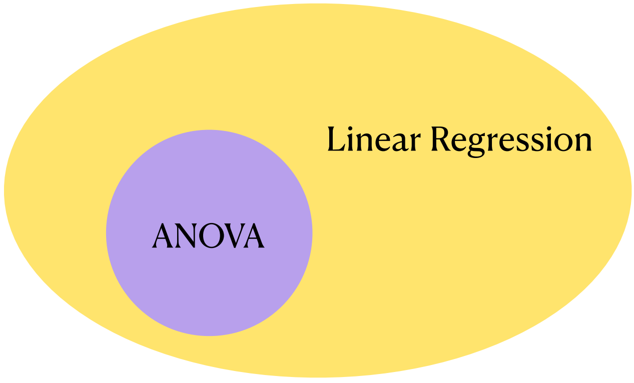
Gender (factor A) i	Bone Development (factor B) j Severely Depressed (B_1)	Moderately Depressed (B_2)	Mildly Depressed (B_3)
Male (A_1)	1.4 (Y_{111})	2.1 (Y_{121})	.7 (Y_{131})
	2.4 (Y_{112})	1.7 (Y_{122})	1.1 (Y_{132})
	2.2 (Y_{113})		
Mean	2.0 ($\bar{Y}_{11\cdot}$)	1.9 ($\bar{Y}_{12\cdot}$)	.9 ($\bar{Y}_{13\cdot}$)
Female (A_2)	2.4 (Y_{211})	2.5 (Y_{221})	.5 (Y_{231})
		1.8 (Y_{222})	.9 (Y_{232})
		2.0 (Y_{223})	1.3 (Y_{233})
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 **<u>linear model</u>** 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 ith treatment effect: $\tau_i = \mu_i - \mu_.$

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

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

$\rightarrow \tau_r = -\tau_1 - \ldots - \tau_{r-1}$

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

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:

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

Dependent variable

independent variable

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

Regression Formulation to Single-Factor Studies

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

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

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

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

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

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

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:

$H_0 : \tau_1 = \ldots = \tau_{r-1} = 0$ $\quad$ $H_a :$ not all $\tau_i' s$ equal to zero

restricted model

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

Full or unrestricted model

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

$$SSE(R) = SSTO$$

$$SSE(F) = SSE$$

Idea:
More parameters included in the model —> better one can fit the data —> 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

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

$$= \frac{\frac{SSTR}{df_{SSTR}}}{\frac{SSE}{df_{SSE}}} = \frac{MSTR}{MSE}$$

ANOVA's test for equality of factor level means!