NRES_798_9_201501

ANOVA models

Model selection in multiple linear regression

$$Y \sim Normal\left(b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2, \sigma^2\right)$$

- E.g. Six primary variables, 15 two-way interactions, 20 three-way interactions

- r^2
  - RSS won't increase with additional terms.
  - What increase in r^2 is sufficient to keep term?
- p-value of slope coefficients?
  - Coefficient value and significance depends on other covariates and order in model.

- Limited number of observations (n)
  - Rule of thumb, don't try to estimate more than n/3 parameters

- Forward selection, Backward selection, Stepwise method
- Objective of analysis is important: hypothesis testing vs. predictive

Model selection in multiple linear regression

$$Y \sim Normal\left(b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2, \sigma^2\right)$$

- Forward selection
  - Add variables until a stop criteria is achieved
  - Criteria normally based on F-ratio and tolerance

$$F - ratio = \frac{explained\ variance}{unexplained\ variance} = \frac{SS_{reg}/1}{RSS/(n-2)}$$

    - F-ratio dependent on both r^2 and number of parameters
    - Add terms until the increase in the F-ratio drops below a threshold.
  - Tolerance
    - Degree of multicollinearity among variables.

$$Tolerance = (1 - r^2)$$

    - If tolerance is too low, the variable is not added to the set (i.e. too much collinearity with other variables)

Model selection in multiple linear regression

$$Y \sim Normal\left(b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2, \sigma^2\right)$$

- Backward elimination
  - Begin with a full model and eliminate variables sequentially
    - Remove variables until there is too large a drop in the F-ratio
    - Remove variables until the increase in the tolerance surpasses a threshold

Model selection in multiple linear regression

$$Y \sim Normal\left(b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2, \sigma^2\right)$$

- Stepwise method
  - Combination of forward and backward selection to evaluate the importance of specific variables
  - Variables added and removed and impact on F-ratio and tolerance used to determine inclusion

- Problems
  - Finding the “best” set of predictor variables very difficult
  - “Optimal” regression model not compared with 2nd, 3rd best models

- Best practices
  - If possible, run all models and evaluate output (only possible if initial set of variables is not too large).
  - Don’t use multiple regression for data mining.
  - Clear objectives will lead to clear model formulation.

ANOVA framework

- Continuous response variable, categorical predictor variables
- Fisher's approach for partitioning Sum of Squares
  - Total variation in data expressed as Sum of Squares

$$SS_y = SS_{reg} + RSS$$

$$ss_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

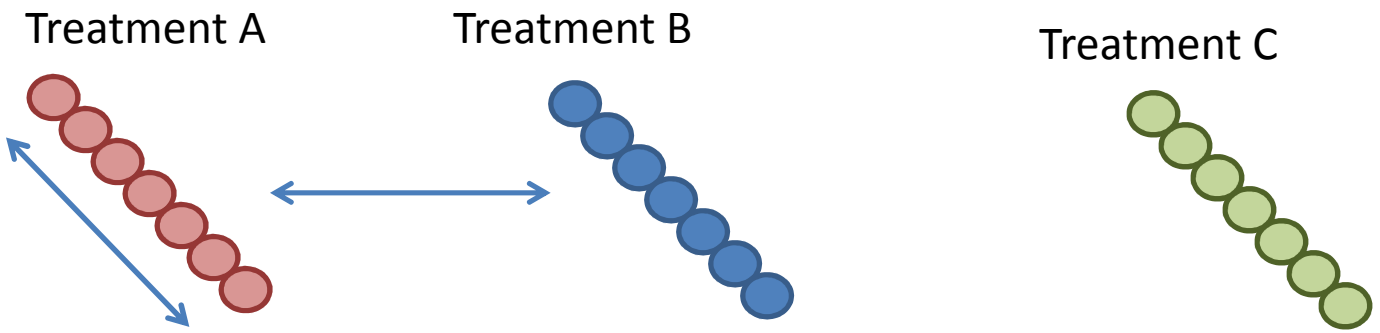
$$RSS = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_i\right)^2$$

One-way ANOVA
3 treatments, a = 3
4 replicates per treatment, n = 4
Total of 12 observations

$$SS_{total} = \sum_{i=1}^{a} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}\right)^2$$

Variance partitioning

- Within and between group variance



Variance partitioning

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}\right)^2$$

- Variation among groups

Treatment group mean

$$SS_{among\ groups} = \sum_{i=1}^{a}\sum_{j=1}^{n}(\bar{Y}_i - \bar{Y})^2 = \sum_{i=1}^{a} n_i(\bar{Y}_i - \bar{Y})^2$$

Grand mean

- Variation within groups

$$SS_{within\ groups} = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_i\right)^2$$

Treatment group mean

$$SS_{total} = SS_{among\ groups} + SS_{within\ groups}$$

Residual sum of squares
Residual variance
Error variation

Hypothesis testing and F statistics

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

$$Y_{ij} = \mu + A_i + \varepsilon_{ij}$$

Grand mean

Coefficient representing additive linear component

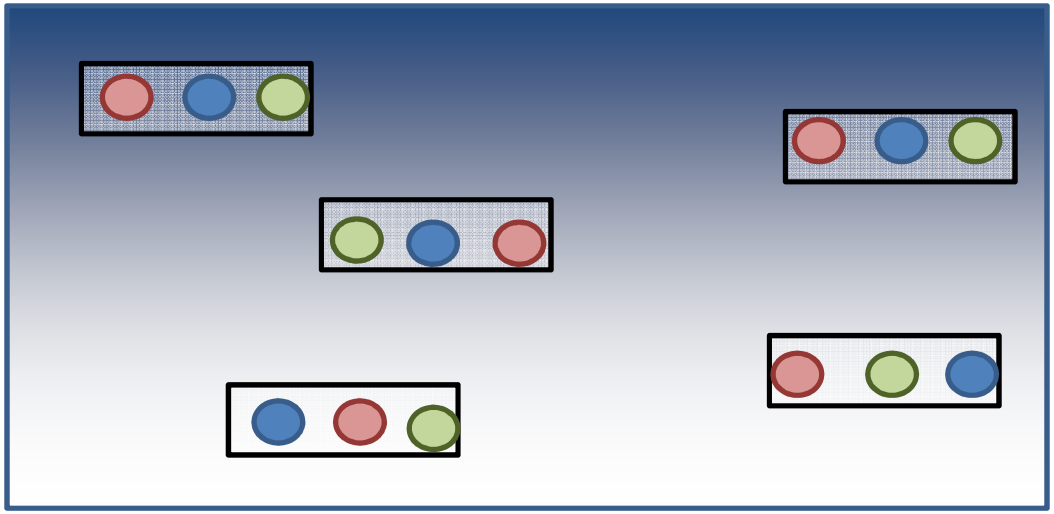
- Null hypothesis $A_i = 0$
- Alternative hypothesis A_i != 0

$$F - ratio = \frac{explained\ variance}{unexplained\ variance} = \frac{MS_{among}}{MS_{within}}$$

ANOVA assumptions

1. Samples are independent and identically distributed
2. Variances are homogeneous among groups
3. Residuals are normally distributed
4. Samples are classified correctly
5. Main effects are additive
   - No interaction between the factors
   - Important in random block and split-plot designs

Random block design



$$Y_{ij} = \mu + A_i + B_j + \varepsilon_{ij}$$

No interaction term

Random block design

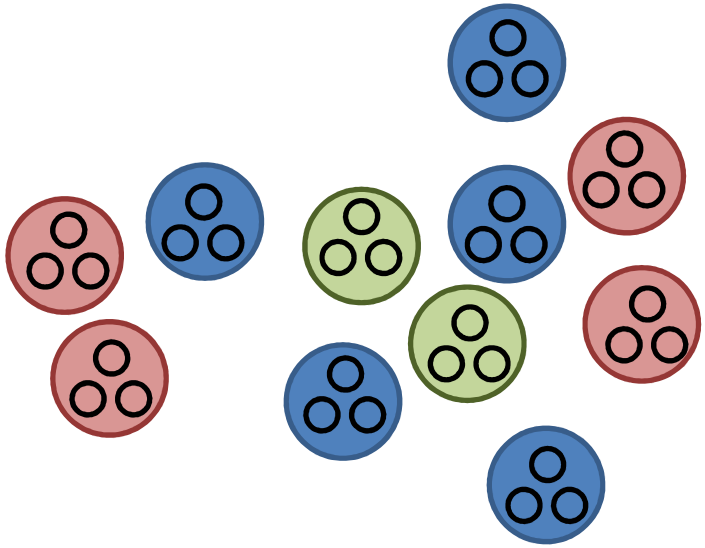
$$SS_{among\ groups} = \sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{Y}_i - \bar{Y})^2 \qquad F - ratio = \frac{MS_{among\ groups}}{MS_{within\ groups}}$$

$$SS_{blocks} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{Y}_j - \bar{Y}\right)^2 \qquad F - ratio = \frac{MS_{blocks}}{MS_{within\ groups}}$$

$$SS_{within\ groups} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(Y_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}\right)^2$$

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\left(Y_{ij} - \bar{Y}\right)^2$$

Nested design



<u>Nesting</u>
Genera
Family

i = 1 to a treatments
j = 1 to b replicates per treatment
k = 1 to n subsamples per replicate

$$Y_{ij} = \mu + A_i + B_{j(i)} + \varepsilon_{ijk}$$

Variation among replicates which are nested within treatments

Nested ANOVA

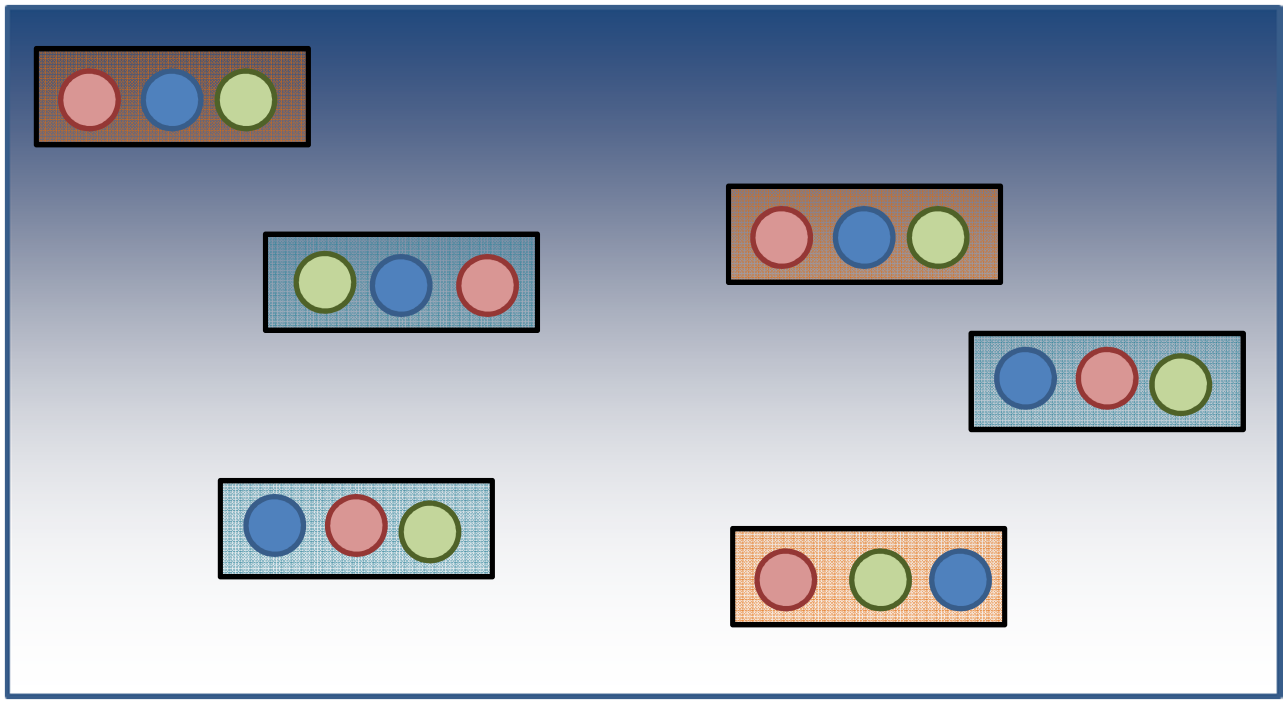
$$SS_{among\ groups} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_i - \bar{Y})^2 \qquad F - ratio = \frac{MS_{among\ groups}}{MS_{among\ replicates\ (groups)}}$$

$$SS_{replicates(groups)} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(\bar{Y}_{j(i)} - \bar{Y}_i\right)^2 \qquad F - ratio = \frac{MS_{among\ replicates\ (groups)}}{MS_{subsamples}}$$

$$SS_{subsamples} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}_{j(i)}\right)^2$$

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}\left(Y_{ijk} - \bar{Y}\right)^2$$

Split plot design



Whole plot factor

Subplot factor

Whole-plot treatment

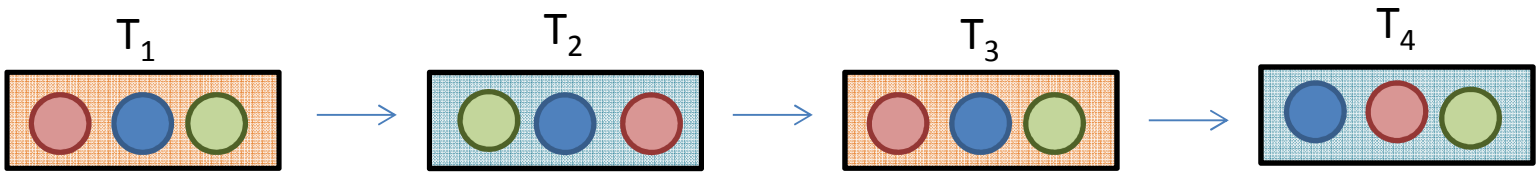
Within plot factor

$$Y_{ijk} = \mu + A_i + B_{j(i)} + C_k + AC_{ik} + CB_{kj(i)}[\varepsilon_{ijk}]$$

Subplots nested within factor A

Error term can't be separated from subplot and within plot variance

Repeated measures



$$Y_{ij} = \mu + A_i + B_j + \varepsilon_{ij}$$

Simple randomized treatments through time (randomized block)

Single treatment per individual measured through time

$$Y_{ijk} = \mu + A_i + B_{j(i)} + C_k + AC_{ik} + CB_{kj(i)}$$

between-subject effects (treatment)

Time factor

Subject by time interaction

Treatment by time interaction

Within-subject effects

Two-way ANOVA

Factor B	Factor A: 1	2	3	4
Aa	5	5	5	5
Bb	5	5	5	5
Cc	5	5	5	5

$$Y_{ij} = \mu + A_i + B_j + AB_{ij} + \varepsilon_{ijk}$$

Two-way ANOVA

$$SS_{factor\ A} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_i - \bar{Y})^2$$

$$F - ratio = \frac{MS_A}{MS_{within\ groups}}$$

$$SS_{factor\ B} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_j - \bar{Y})^2$$

$$F - ratio = \frac{MS_B}{MS_{within\ groups}}$$

$$SS_{interaction} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(\bar{Y}_{ji} - \bar{Y}_i - \bar{Y}_j + \bar{Y})^2$$

$$F - ratio = \frac{MS_{AB}}{MS_{within\ groups}}$$

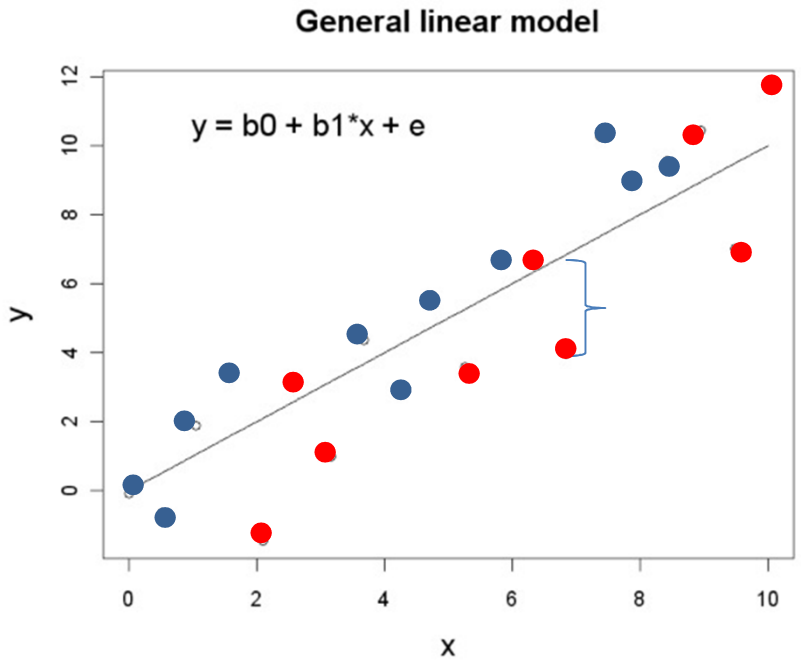
$$SS_{within\ groups} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ji})^2$$

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y})^2$$

ANCOVA

ANOVA $Y_{ij} = \mu + A_i + \varepsilon_{ij}$

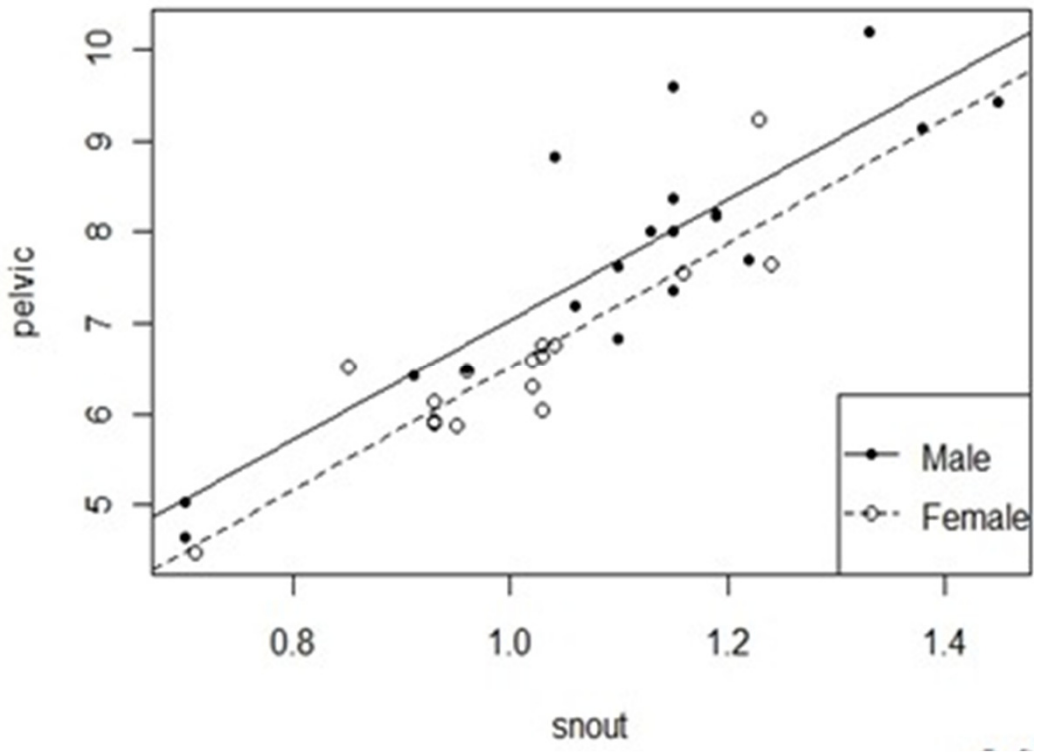
Environmental variation



- ANCOVA
  - Analysis performed on the residuals from a regression of the response variable on the covariate

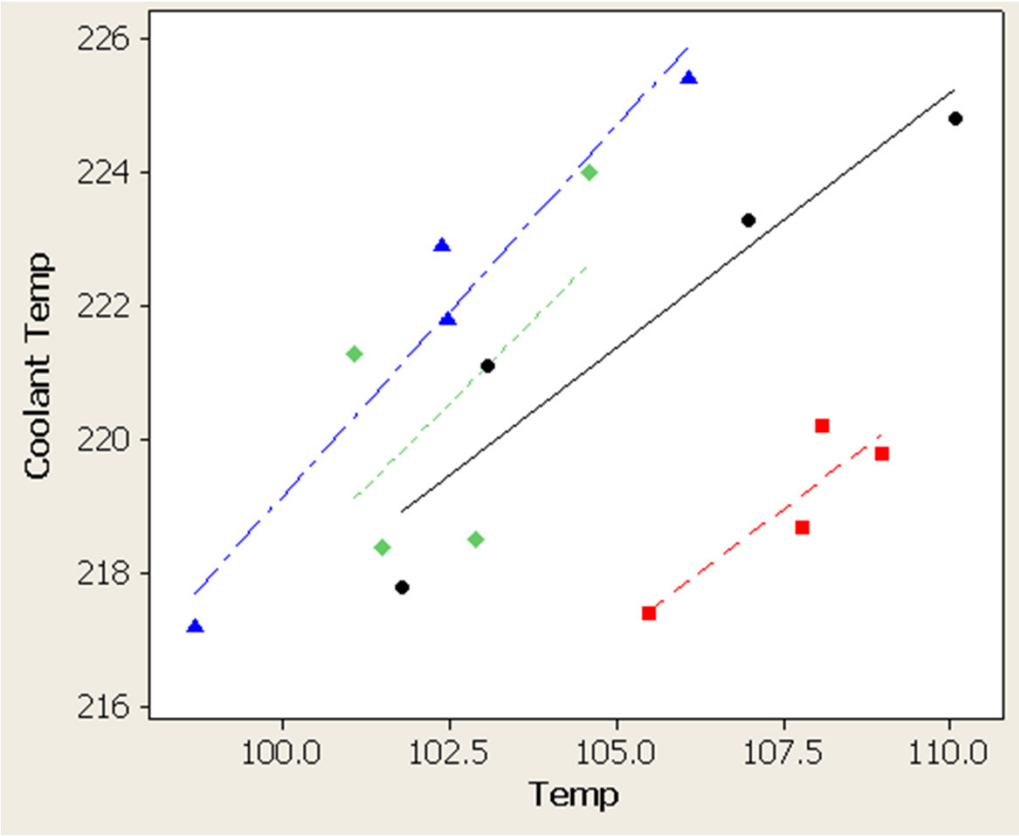
ANCOVA

$$Y_{ij} = \mu + A_i + B_C\left(X_{ij} - \bar{X}_i\right) + \varepsilon_{ij}$$



ANCOVA

$$Y_{ij} = \mu + A_i + B_C(X_{ij} - \bar{X}_i) + \varepsilon_{ij}$$



$$Y_{ij} = \mu + A_i + B_i(X_{ij} - \bar{X}_i) + \varepsilon_{ij}$$

Robustness of ANOVA design

1. Samples are independent and identically distributed
2. Variances are homogeneous among groups
3. Residuals are normally distributed
4. Samples are classified correctly
5. Main effects are additive