TTK31 - Design of Experiments (DoE), metamodelling and Quality by Design (QbD) Autumn 2021

Big Data Cybernetics Gang

Lecture overview

1. DoE: Introduction and motivation
2. **ANalysis Of VAriance (ANOVA)**
3. Factorial designs
4. Fractional factorial designs
5. Response surface designs
6. Optimal designs
7. Metamodelling
8. Combining DoE with multivariate analysis/machine learning
9. QbD – PAT
10. Practical examples of DoE related to cybernetics

Reminder: Reference group - VERY IMPORTANT

- at least 3 students
- will do 4 meetings (1 after the exam)
- shall represent the whole class $\implies$ you will have meetings among yourselves too
- shall lead to a <u>referansegrupperapport</u> containing suggestions for improvements

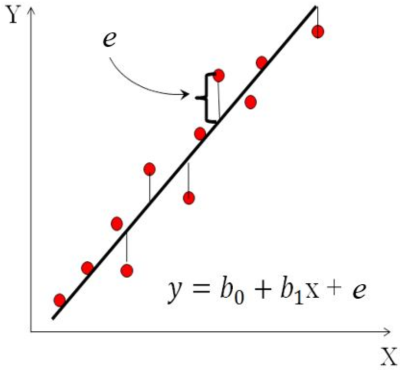
Multiple Linear Regression (MLR) and ANOVA

ANalysis of VAriance (ANOVA)

- ANOVA is the most frequently used way to analyse results from design of experiments
- The main purpose is to estimate the variance in the responses due to the various model terms and assess if the model terms are significant
- Extensions:
    - ANCOVA (additional variables; covariates)
    - MANOVA (simultaneous analysis of several responses)
    - MANCOVA (several responses and covariates)

Linear Regression - the univariate case

- Fit a straight line to the data
- The parameters $\mathbf{b_0}$ and $\mathbf{b_1}$ need to be estimated
- The aim of least squares is to minimize the squared sum of the error terms, $\mathbf{e}$
- Thus, the assumption is that there are no errors in $\mathbf{X}$
- MLR require more objects than variables (influences the design matrix as it depends on the choice of model complexity)

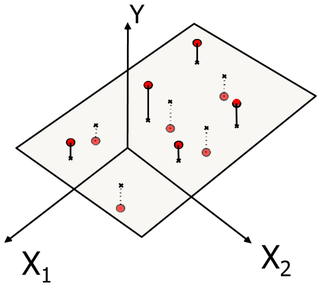


Multiple Linear Regression (MLR) - general case

The model equation, which relates a response variable to several predictors by means of regression coefficients, has the following structure:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + e$$

- Least Squares criterion: the plane should lie where it minimizes the sum of squares of all residuals.
- The method of choice for orthogonal experimental designs
- The ANOVA table is calculated from the regression coefficients in most implementations; in the earlier days by means of square sums (programming it yourself should take 10-15 minutes :-)



Modelling statistics and diagnostics

Modelling statistics

Residual sum of squares

$SS_{residuals} = \sum_{i=1}^{n} e_i^2$

R-squared: The amount of variance explained by the model

$$R^2 = 1 - \frac{SS_{residuals}}{SS_{residuals} + SS_{model}}$$

Adjusted R-squared:R-squared adjusted for the number of terms in the model

$$R^2 = 1 - \left(\frac{SS_{residuals}}{df_{residuals}}\right) / \left(\frac{SS_{residuals} + SS_{model}}{df_{residuals} + SS_{model}}\right)$$

Model diagnostics

- Before concluding on the ANOVA results, one needs to investigate various model diagnostics
  - The structure of the residuals
  - The impact on the model for the individual samples
  - The goodness of the model
- Some statistical figures of merit
  - Leverage (hat matrix)
  - Cook's distance

Model diagnostics - Leverage

The hat matrix $\mathbf{H}$ is given by:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}$$

Leverage for sample *i*:

$h_i =$ diagonal element of $\mathbf{H}$:

$$h_i = \mathbf{x_i}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x_i^T}$$

Leverage is a value between 0 and 1

Model diagnostics - residuals

Estimate of the standard deviation, MSE:

$$\hat{\sigma} = \sqrt{\sum_{i}^{I} e_{i=1}^{2} / (I - K - 1)}$$

Internally Studentized Residual: The residual divided by the estimated standard deviation of that residual. It measures the number of standard deviations separating the actual and predicted values.

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

More modelling statistics

PRESS: Predicted Residual Error Sum of Squares. The error when predicting a sample with a model of which the sample was not included

$$e_{-i} = y_i - \hat{y_{-i}} = \frac{e_{-i}}{1 - h_{ii}}$$

$$PRESS = \sum_{i=1}^{n} e_{-i}^2$$

where

h_{ii} is the leverage of sample *i*

$$\text{RMSE (Root Mean Square)} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{I}\sum_{i=1}^{I} e_i^2}$$

Model diagnostics - Cook's distance

Cook's distance is a diagnostic that takes into account the residual as well as the leverage of one sample

It represents the change in the regression line when the sample i is removed

Cook's distance $= D_i = \frac{r_i^2}{k+1}\left(\frac{h_i}{(1-h_i)}\right)$

The F-distribution and relationship to the t-test

- F-tests are named after Sir Ronald Fisher, who developed the theory in the 1920's
- The F-statistic is simply a ratio of two variances estimated as mean squares corrected for degrees of freedom (DF)
- Rule of thumb: If the variance due to changing a design factor is three times the noise/error, it is most likely not due to chance
- If you have only two groups/factor levels, the F-test statistic is the square of the t-test statistic, and the F-test is equivalent to the two-sided t-test.
- The calculated test statistic F_0 is compared to an F-table for a specified number of degrees of freedom. The form of the test statistic is as follows: $F_{\alpha,n_{1-1},n_{2-1}}$
- α is the significance level that you will decide upon *a priori*

ANalysis Of VAriance (ANOVA)

- ANOVA separates data into contributions from structure and noise

- Data = Structure + Noise
- $SS_{Total} = SS_{Model} + SS_{residuals}$
- Total variation = Modelled + Not modelled

$$\sum_{i=1}^{I}(y_i - \bar{y_i})^2 = \sum_{i=1}^{I}(y_i - \hat{y_i})^2 + \sum_{i=1}^{I}(\hat{y_i} - \bar{y_i})^2$$

ANOVA output (1/2)

Summary-section:

- Model (SS_{Model}):
  - Contribution form all terms in the model
  - Degrees of Freedom (DF) is given by the number of estimated model parameters
- Error ($SS_{residuals}$):
  - Non-modelled variation or noise
  - DF given by number of (runs - number of terms - 1)
- Significance of model is estimated from

$$\text{F-ratio} = \frac{\text{MS}_{\text{Model}}}{\text{MS}_{\text{residuals}}}$$

ANOVA output (2/2)

- Variables-section:
  - The significance of each model parameter is estimated
- Model check selection
  - Sums the contribution from linear terms, interaction terms etc.
- Lack-of-fit section
  - Total error may be divided into
    - Pure error: Spread between replicates
    - Lack-of-fit: Modelled values vs. Mean of replicates

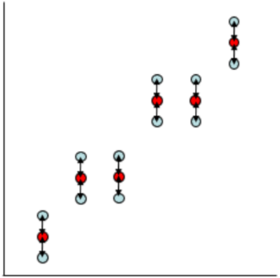
Lack of fit

Testing Lack of fit is important for evaluating the goodness of the model

$SS_{residuals} = SS_{pure\ error} + SS_{lack\ of\ fit}$

$SS_{pure\ error}$ = SS of the replicates about their means

$SS_{lack\ of\ fit}$ = SS of the means about the fitted model.



$$F = \frac{MS_{lack\ of\ fit}}{MS_{pure\ error}}$$

Is the variation about the model greater than what is expected given the variation of the replicates about their means?

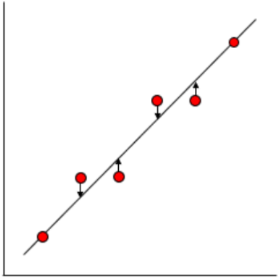
Residual distance

Lack of fit is compared to the residual distance to the model

$SS_{residuals} = SS_{pure\ error} + SS_{lack\ of\ fit}$

$SS_{pure\ error}$ = SS of the replicates about their means

$SS_{lack\ of\ fit}$ = SS of the means about the fitted model.



$$F = \frac{MS_{lack\ of\ fit}}{MS_{pure\ error}}$$

Is the variation about the model greater than what is expected given the variation of the replicates about their means?

ANOVA table

The ANOVA table for the overall model has the following structure:

Table: The structure of the ANOVA table for the overall model

Source	**SS**	**df**	**MS**	**F-ratio**	p-value
Model	SS_{Model}	k	MSR $= SS_{Model}/$(k)	MSR/MSE	*p*
Error	SS_{Error}	l-k-1	MSE $= SS_{Residual}/$(l-k-1)		
Total	SS_{Total}	l-1	MST $= SS_{Total}/$(l-1)		

ANOVA for the individual model terms

The ANOVA table for the individual model terms has the following structure, here shown for a model with two variables at two levels:

Table: The structure of the ANOVA table for the individual model terms

Source	**SS**	**df**	**MS**	**F-ratio**	p-value
Intercept	$SS_{Intercept}$	1	MS	MS/MSE	*p*
Variable1	$SS_{Variable1}$	1	MS	MS/MSE	*p*
Variable2	$SS_{Variable2}$	1	MS	MS/MSE	*p*

ANOVA in case of unbalanced designs and non-orthogonal designs

- When the design is orthogonal, the way the square sums is estimated will not alter the ANOVA table
- In other cases there is no "truth" in how to estimate the square sums
- In short, the options are:
  - Type I: Estimate the square sums sequentially: First assign a maximum of variation to variable A; in the remaining variation, assign the maximum of variation to variable B.
  - Type II: This type tests for each main effect after the other main effect SS(A|B), SS(B|A). The common option if you are mostly interested in the main effects.
  - Type III: Estimate square sums while taking into account all other model terms; For models with interactions terms

A small example to illustrate MLR and ANOVA

- House prices in the Boston area, 434 samples
- Model: Median value of property = f(No. of rooms, age)
- Demo in Design-Expert®
  - Look at raw data
  - ANOVA
  - Diagnostics plots
  - Model statistics

Multiple Linear Regression details

MLR details - I

Estimation of regression coefficients $\mathbf{b}$

$$\hat{\mathbf{b}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{1}$$

NB! This requires full rank in $\mathbf{X}$

The fitted values of $\mathbf{y}$ (prediction from the calibration):

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{b}} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{2}$$

The Y-residuals, e_i are given by

$$e_i = y_i - \hat{y}_i \tag{3}$$

MLR details - II

The error standard deviation is estimated as

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{I} e_i^2/(I - K - 1)} \tag{4}$$

The variance of b_0:

$$\hat{\sigma}_{b_0} = \hat{\sigma}\left[\frac{1}{I} + \frac{\bar{\mathbf{X}}^{\mathbf{2}}}{\mathbf{X}^{\mathbf{T}}\mathbf{X}}\right] \tag{5}$$

The variances of $b_1, ..., b_K$:

$$\hat{\sigma}_b = \hat{\sigma}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1} \tag{6}$$

MLR details - III

The t-statistic for the beta coefficients is:

$$t = \frac{\hat{b}}{\hat{b}_\sigma} \tag{7}$$

The critical t-value is given by the t-distribution with (I-K-1) degrees of freedom.
The confidence interval for b:

$$\hat{b} \pm t_{\alpha/2}\hat{\sigma}_b \tag{8}$$