Nonlinear regression
Mediation vs. Confounding
Experimental design

Anton Rask Lundborg
arl@math.ku.dk

Copenhagen Causality Lab
Department of Mathematical Sciences

January 3, 2024

Lecture outline

- Non-linear regression.
- Causal diagrams:
  - Mediation and confounding.
  - Path analysis.
- Elements of experimental design:
  - Replication, Randomization, Blocking, Covariates, Multifactorial.
  - Generation of optimal designs.

Summary (Day 1)

- Statistics answers four important questions:
  1. Is there an effect? (falsification of null hypotheses, p-values)
  2. Where is the effect? (p-values from post hoc analyses)
  3. What is the effect? (confidence and prediction intervals)
  4. Can the conclusions be trusted? (model validation)

- We do model-based frequentist statistics: Interpretation of p-values and confidence intervals via the meta-experiment.

- Tidy datasets consist of variables (columns) and observations (rows).

Can the conclusions be trusted? — Validation of linear normal models

- **Do residuals have mean=0?** Plot of residuals vs predicted values.
- **Do residuals have same variance?** Plot of standardized residuals vs predicted values.
  - Often see that variance increases with predicted values (forms a "trumpet shape"). Usually solved by a log-transformation.
- **Are residuals normal distributed?** Normal quantile plot of standardized residuals.
  - Banana shape indicates need for log-transformation.
- **Are residuals independent?** Not validated formally. Instead, use knowledge about design of experiment. See also course Day 5.
- **Are there any outliers?** Plot of standardized residuals vs leverages. Critical lines in terms of Cook's distances (D=0.5, D=1.0).
  - Generally it is not advisable to remove observations!
  - Determining robustness of results when excluding some observations can be tried.

Non-linear regression

Non-linear regression – possible solution when residuals are not mean zero

- Continuous response y (the dependent variable).
- Continuous covariates $x_1, \ldots, x_K$ (the independent variables).
- Assume there exists a parameter θ and a function f_θ such that for every observation i,

$$y_i = f_\theta(x_{1i}, \ldots, x_{Ki}) + \epsilon_i, \qquad (\epsilon_i)_{i=1}^n \text{ independent } \mathcal{N}(0, \sigma^2)$$

- Parameter estimate $\hat{\theta}$ minimizes the sum of squared errors

$$\sum_{i=1}^{N} \left| y_i - f_\theta(x_{1i}, \ldots, x_{Ki}) \right|^2$$

- Parameter σ estimated by the Root-Mean-Squared-Error (RMSE)

$$\hat{\sigma} = \sqrt{\frac{1}{N-p} \sum_{i=1}^{N} \left| y_i - f_{\hat{\theta}}(x_{1i}, \ldots, x_{Ki}) \right|^2}, \qquad p = \dim \theta$$

Some model examples

- One covariate x and two parameters α, β:

$$y = \alpha + \beta x \qquad \text{(the straight line)}$$

$$y = \frac{1}{\alpha + \beta x} \qquad \text{(an inverse line)}$$

$$y = \alpha \cdot e^{\beta x} \qquad \text{(exponential function)}$$

$$y = \alpha \cdot x^{\beta} \qquad \text{(power function)}$$

- One covariate x and four parameters α, β, γ, δ:

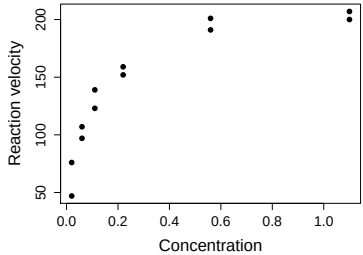
$$y = \alpha + \beta x + \gamma x^2 + \delta x^3 \qquad \text{(cubic regression)}$$

- The model is called linear if it is a linear function in the model parameters.
  - **Quiz:** Which of the above models are linear?
  - **Quiz:** Which of the above models can be made linear by using transformations?

Data example: Puromycin

Reaction velocity (y) as a function of concentration (x)

concentration	reaction
0.02	76
0.02	47
0.06	97
0.06	107
0.11	123
0.11	139
0.22	159
0.22	152
0.56	191
0.56	201
1.10	207
1.10	200



Michaelis-Menten model with parameters α (max velocity) and β (concentration at max/2):

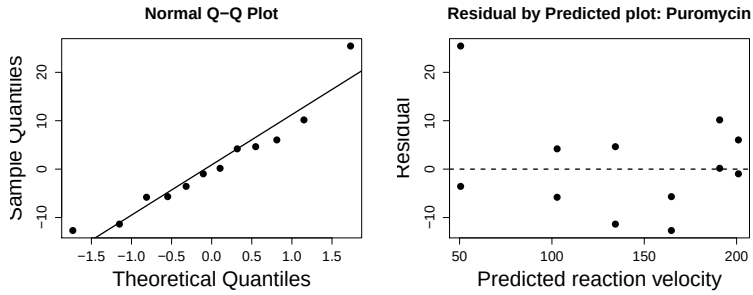
$$y = \frac{\alpha x}{\beta + x}$$

Puromycin: Non-linear regression (I) – see `puromycin.R`

```
# Read data from text file, and make plot
puromycin <- read.delim("puromycin.txt")
plot(reaction ~ concentration, data = puromycin)

# Fit non-linear regression: Guesses taken from plot
m1 <- nls(reaction ~ a * concentration / (b + concentration),
          start = list(a = 200, b = 0.1), data = puromycin)

# Can the conclusions be trusted?
# Normal quantile plot:
qqnorm(residuals(m1))
abline(mean(residuals(m1)), sd(residuals(m1)))

# Residual plot:
plot(predict(m1), residuals(m1))
abline(0, 0, lty=2)
```

Puromycin: Non-linear regression (II) – see `puromycin.R`

Neither the normal quantile plot nor the residual plot are too good (largely due to 1 observation at the lowest concentration):

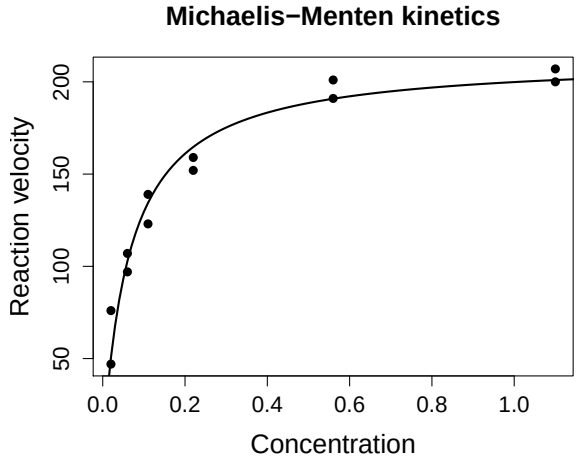


To be on the safe side we do a Lack-of-Fit test, which gives $p = 0.4468$:

```
m0 <- lm(reaction ~ factor(concentration), data=puromycin)
plot(m0)
anova(m1, m0)
```

Puromycin: Fitted curve ($\hat{\alpha} = 212.68$, $\hat{\beta} = 0.0641$)

Neither confidence nor prediction intervals available in `predict.nls()`.



Exercise 4.2 revisited: World running records – see `wr2011_nls.R`

- World records in running are (rather) well-modelled on the log vs. log scale if we allow for a change-point between short and long running distances:

$$\log(\text{time}) \approx \alpha(\text{sex}) + \beta \cdot \log(\text{distance}) + \gamma \cdot \log\left(\frac{\max(\delta, \text{distance})}{\delta}\right)$$

- This model is linear in the parameters $\alpha(\text{men}), \alpha(\text{women}), \beta, \gamma$ and non-linear in the parameter δ. Here δ encodes the change-point.

- Non-linear regression gives the estimates:

$$\hat{\alpha}(\text{men}) = -3.4480, \qquad \hat{\alpha}(\text{women}) = -3.3365,$$
$$\hat{\beta} = 1.2059, \qquad \hat{\gamma} = -0.1379,$$
$$\hat{\delta} = 1212.9478.$$

Please note that `nls()` does not always find the best model fit!

Checkpoint

- Questions?
- After the break we discuss confounding and mediation. This is done using causal diagrams.
  - Causal inference is a hot topic in contemporary methodological research in statistics.
  - We then move on to discuss design of experiments.

Time for a break!

Mediation vs. Confounding

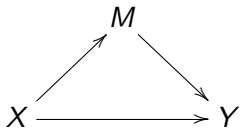
Causal diagrams

- Causal diagrams display the causal relations that the scientist (not the statistician) believes/knows to be present in advance.
  - It is difficult (although possible in some situations) to infer this from observational data.
  - Relations can be established by randomized trials.
- In multilinear regressions (i.e. many explanatory variables) it is important to understand the role of the explanatory variables.
  - Mediators or Confounders?
- When the joint distribution is normal (as discussed today), many things can be done:
  - Correcting for confounders allow for estimation of causal effects.
  - Instrumental variables to deal with non-observed confounders.
  - Path analysis to decompose mediation.
- Similar issues also arise for non-normal responses. Here the mathematics is more difficult, and much of this is ongoing research.

Causal diagrams: Mediation vs. Confounding

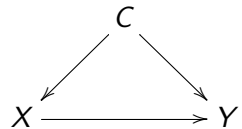
Which variables should be included, and what is the interpretation?

Effect of X mediated through M



- Regression of Y on X gives the total effect.
- Regression of Y on (X, M) gives the direct effect (of X).
- total effect = direct effect + indirect effect.

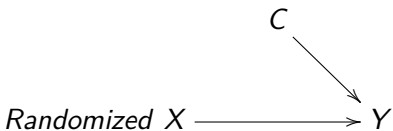
Effect of X confounded by C



- Spurious effects avoided by including the confounders.
- $\Longrightarrow$ effect of X quantified via coefficient on X in the regression of Y on (X, C).

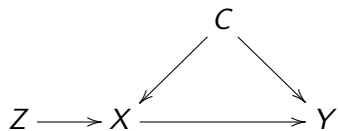
Causal effects with unobserved confounders – two potential solutions

Randomized Controlled Trial



- Relation between confounder C and regressor X is broken by the randomization.
- Effect of X is now causal.
- If confounder C is observed – could be included to reduce variation.

Instrumental variable Z

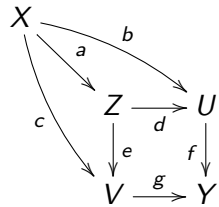


Two-step procedure:

1. Regress X on Z to find predicted value $\hat{X} = \hat{E}[X|Z]$.
2. Regress Y on $\hat{X}$. Coefficient gives causal effect of X.

Removes effect of confounder C. But procedure can have low power!

Path analysis: An arbitrary example with 3 mediators



Solved by 4 regressions (from "right" to "left"):

$$
\begin{aligned}
Y - \mu_Y &= f \cdot (U - \mu_U) + g \cdot (V - \mu_V) + \text{error} \\
V - \mu_V &= c \cdot (X - \mu_X) + e \cdot (Z - \mu_Z) + \text{error} \\
U - \mu_U &= b \cdot (X - \mu_X) + d \cdot (Z - \mu_Z) + \text{error} \\
Z - \mu_Z &= a \cdot (X - \mu_X) + \text{error}
\end{aligned}
$$

Combined these gives a decomposition of the total effect of X on Y:

$$
Y - \mu_Y = \left(f \cdot (b + d \cdot a) + g \cdot (c + e \cdot a)\right) \cdot (X - \mu_X) + \text{error}
$$

Design of Experiments

Design of experiments – Important components

- Replications
- Randomization
- Blocking
- Covariates
- Multifactorial design

Design of experiments – Replications

- Necessary for doing a statistical analysis.
- Needed to estimate size of random variation.
- The only way to know about reproducibility and hence to assess treatment effect.
- More replicates lead to higher precision.

Rule of thumb

Required sample size grows quadratically with required precision.

Design of experiments – Randomization

- Random allocation of experimental units to treatments.
- Mitigates systematic errors (bias), e.g. from confounding.
- Must be carried out by true randomness, e.g. rolling dice, random numbers from computer, atmospheric noise.
- Balancing: Rearrangement making groups alike. May invalidate statistical analysis and should be avoided.
  - Example: Switching animals between groups to make groups similar according to initial weight. This may invalidate an ANOVA since the variation within groups becomes too large.
  - However, see next slide on blocking.

Design of experiments – Blocking

- Grouping of experimental units into homogeneous groups (called blocks).
- Randomization: Units allocated randomly to treatments within block.
- Model should include main effects of used blocks (there might be more than one). Possibly as random effects (see Day 5).
- Reduces residual variation $\Longrightarrow$ Increases precision and power.
- Examples:
  - different areas in a field
  - replication, e.g. day of experiment
  - experimental unit in cross-over designs

Design of experiments – Covariates

- Continuous measure on each unit with possible relation to response.
- Reduces residual variation $\Longrightarrow$ Increases precision and power.
  - See exercise 6.3 for an example of this.
- Should (ideally) not be associated with the treatment
  - If influenced by the treatment: mediation!
  - If influencing the treatment: confounding!
- Several covariates may be used simultaneously in the model.

Design of experiments – Multifatorial design

- Use multifactorial designs — not one factor at a time.
- More information $\Longrightarrow$ more efficient.
- Interactions can be investigated.
  - Analysis done similar to the 2-way ANOVA but now we may also have third or higher order interactions.
  - Often preferable only to include 2-way interactions unless the application calls for more.

Checkpoint

- Questions?
- After the break we continue to discuss design of experiments including computation of power and required sample sizes.

Time for a break!

Example of experimental design: Half-fraction factorial design

Suppose the following 5 factors are believed to influence the percentage of chemicals that respond in a reactor:

- The feed rate of chemicals (FeedRate), 10 or 15 liters per minute.
- The percentage of the catalyst (Catalyst), 1% or 2%.
- The agitation rate of the reactor (AgitRate), 100 or 120 revolutions per minute.
- The temperature (Temp), 140 or 180 degrees Celsius.
- The concentration (Conc), 3% or 6%.

A complete factorial design of 5 factors on 2 levels each requires $2^5 = 32$ experimental units. Suppose you only can afford 16 experimental units and want to estimate all main effects and all two-factor interactions.

Then you are looking for a so-called 2^{5-1}_V design (see Fractional Factorial Design on Wikipedia).

Design Of Experiments in R

- One of the general packages in R for DOE is called `AlgDesign`.
- From a larger collection of experimental units, e.g. the full factorial design, an optimal subset of a prespecified size is selected.
- An often used criterion is the so-called D-optimality.
- You have to do the randomization afterwards "by hand", e.g. using `sample()`.

```
# Create full factorial design for 5 factors on 2 levels each
full_factorial <- gen.factorial(
  levels = 2, nVars = 5,
  varNames = c("FeedRate", "Catalyst", "AgitRate", "Temp", "Conc")
)

# Create a design with 16 units
# for main effects and 2-way interactions between 5 factors on 2 levels
optFederov(~ (FeedRate + Catalyst + AgitRate + Temp + Conc)^2,
  data = full_factorial, nTrials = 16
)$design
```

Sample size and power computations

What is the power to detect differences?

- Designs generated by `AlgDesign`, say, are optimal in some mathematical sense. But how can we know if they actually have sufficient power to answer our scientific question?
- One possibility is to insert simulated data for the response variable and do the statistical analysis. Typically, the response variable is simulated using one of the following two methods:
  1. Using the systematic differences and variances we believe (e.g. from the literature) to be present.
  2. Using the systematic differences we want to be able to detect. In this case it is often convenient to measure in scales of the standard deviation, which amounts to setting the variance to 1.
- This is repeated 10,000 times, say, in order to estimate the power by the fraction of times we observe a significant effect.

Classical sample size computations

These methods are only available for the simplest settings:

- For *t*-tests, 1-way ANOVA and comparisons of proportions the needed sample size may be computed without using simulations:
  - `power.t.test()`, `power.anova.test()`, `power.prop.test()`.
  - A few more methods available in packages `pwr` and `LabApplStat`.

- The following numbers must be specified:
  - Used significance level, typically $\alpha = 0.05$ but remember the perspective from Sterne & Smith, namely to consider a lower significance level.
  - The standard deviation in the population. This number must be known from previous experiments or from the literature.
  - The systematic difference between the treatment groups, possibly the least relevant difference to detect.
  - The desired power, i.e. the chance that you will find a significant effect.

- Alternatively, you may obtain the power as a function of the sample size.

Investigation of experimental designs

Example of a more complicated design

ACTA PHYSIOLOGICA

Acta Physiol 2014, **210**, 84–98

Programming of glucose–insulin homoeostasis: long-term consequences of pre-natal versus early post-natal nutrition insults. Evidence from a sheep model

A. H. Kongsted,[1] M. P. Tygesen,[2] S. V. Husted,[1] M. H. Oliver,[3] A. Tolver,[4] V. G. Christensen,[1] J. H. Nielsen[5] and M. O. Nielsen[1]

1 Department of Veterinary Clinical and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark

2 Cook Medical Europe APS, Bjaeverskov, Denmark

3 Ngapouri Farm Research Laboratory, Liggins Institute, University of Auckland, Auckland, New Zealand

4 Department of Basic Sciences and Environment, Faculty of Science, University of Copenhagen, Frederiksberg, Denmark

5 Department of Biomedical Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, København N, Denmark

Nutrition and response to glucose challenge

20 sheep pregnant with twins were randomized to two groups:

1. 10 sheep were given low energy feed (`prenatal=LOW`)
2. 10 sheep were given normal feed (`prenatal=NORM`)

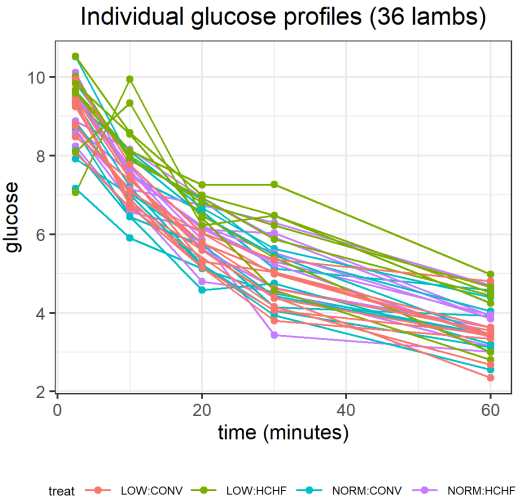
After parturition the 20 twin lambs were separated:

1. one twin was given normal feed (`postnatal=CONV`)
2. one twin was given high energy feed (`postnatal=HCHF`)

36 out of the 40 lambs survived at age 6 months.

At age 6 months the glucose content in serum was measured at 2.5, 10, 20, 30, and 60 minutes after an injection with a large dose of glucose.

Visualization of glucose data



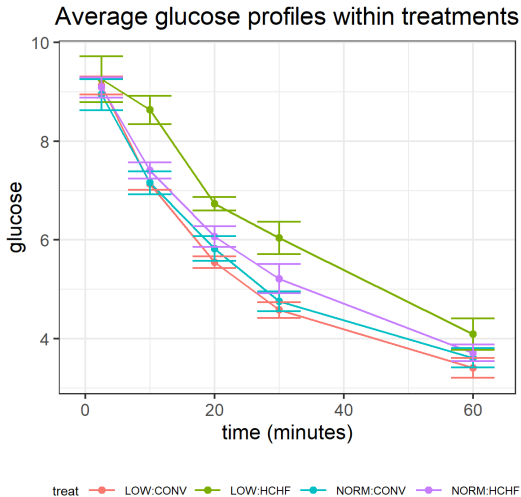


Table of Variables

The variable **names**, their **type** (nominal, ordinal, interval, ratio), **range** (levels for categorical and range for continuous variables), and **usage** (fixed effect, random effect, response).

Table-of-Variables for glucose example:

Variable	Type	Range	Usage
lamb†	Nominal	36 levels	Random effect
sheep	Nominal	20 levels	Random effect
prenatal	Nominal	LOW, NORM	Fixed effect
postnatal	Nominal	CONV, HCHF	Fixed effect
time†	Ordinal	$2.5 < 10 < 20 < 30 < 60$	Fixed effect
	Continuous	[2.5 ; 60]	Fixed effect
glucose	Continuous	[2.347 ; 10.522]	Response

† With many repeated measurements (here 5 per lamb) it is recommended to model temporal correlation, in which case lamb and time are also used as subject ID and serial correlation, respectively. We will skip this here (as this is not implemented in the `LabApplStat`-package).

Design Diagram

Two glucose measurements went wrong $\Longrightarrow N = 36 \cdot 5 - 2 = 178$

Table of Variables:

Variable	Type	Range	Usage
lamb	Nominal	36 levels	Random effect
sheep	Nominal	20 levels	Random effect
prenatal	Nominal	LOW, NORM	Fixed effect
postnatal	Nominal	CONV, HCHF	Fixed effect
time	Ordinal	$2.5 < 10 < 20 < 30 < 60$	Fixed effect
glucose	Continuous	[2.347 ; 10.522]	Response

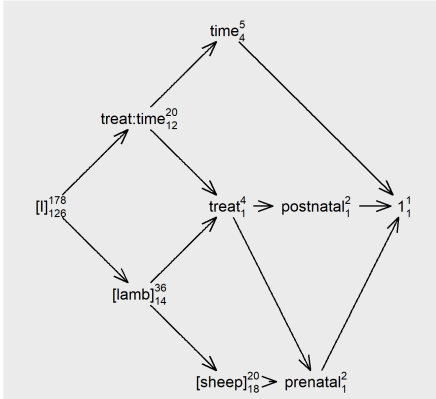
To investigate synergistic / antagonistic effect between prenatal and postnatal we add the interaction

$$\text{treat} = \text{prenatal:postnatal}$$

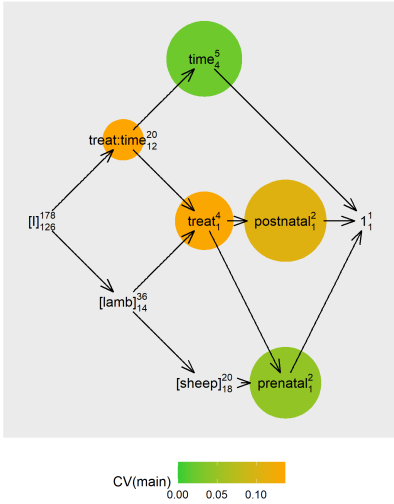
that has 4 levels:

LOW:CONV, LOW:HCHF,
NORM:CONV, NORM:HCHF

Design Diagram for Lamb glucose experiment



Visualization of design information



If [I] is the only random effect, then we do not use the response to compute the design diagram with CV!

- Area of circle proportional to average information about parameters:
  - Larger information $\Longrightarrow$ more narrow confidence intervals + higher power.
- Colors visualize coefficient of variation for the information within each fixed effects.

Some conclusions:

- prenatal is better estimated than postnatal.
- Estimation of treat somewhat unbalanced.
- Good estimation of time effect.

Checkpoint

- Questions?
- After the break we discuss some classical designs.

Time for a break!

Some classical designs

Some common multi-factorial designs

- Full Factorial Design
  - Allows for estimation of all interactions, but requires many experimental runs.
  - Complete Randomized Block Design, where a full factorial design is done within each "batch".
- Fractional Factorial Design
  - Cheaper than full factorial design while retaining control of which interactions are estimable.
  - Entanglement of effects (known as aliasing)
- Plackett-Burman Design
  - Estimation of up to $N - 1$ dichotomous (yes/no) effects using N experimental runs. Exists if N is a multiple of 4.
  - May be constructed using the `pb()` function from the `FrF2`-package.
  - Often used for screening of many effects using few resources.
  - Trade-off: more repetitions or more factors?

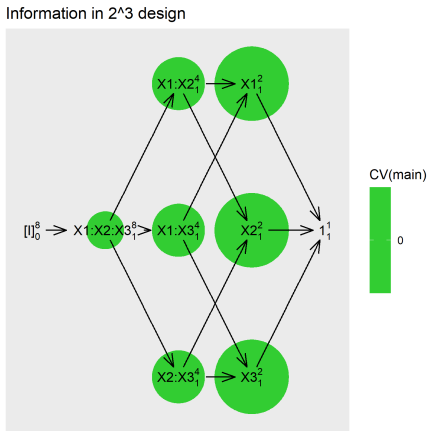
Full Factorial Design: Try all combinations

Remember to shuffle the order of the experimental runs!

3 dichotomous (yes/no) factors:
$N = 2^3 = 8$. For instance,

$X1 : -/+ =$ low/high temperature,
$X2 : -/+ =$ without/with enzyme,
$X3 : -/+ = 5/10$ minutes reaction.

	X1	X2	X3
1	+	+	+
2	−	+	+
3	+	−	+
4	−	−	+
5	+	+	−
6	−	+	−
7	+	−	−
8	−	−	−



Complete Randomized Block Design

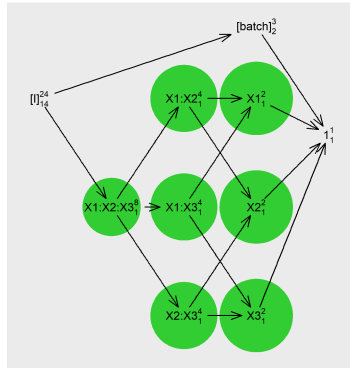
	X1	X2	X3	batch
1	−	+	+	1
2	−	+	−	1
3	+	+	+	1
4	+	−	+	1
5	+	+	−	1
6	−	−	−	1
7	−	−	+	1
8	+	−	−	1
9	−	+	−	2
10	−	−	−	2
11	+	−	−	2
12	+	+	−	2
13	−	+	+	2
14	+	−	+	2
15	−	−	+	2
16	+	+	+	2
17	−	+	+	3
18	+	+	−	3
19	+	+	+	3
20	+	−	+	3
21	−	−	−	3
22	−	+	−	3
23	−	−	+	3
24	+	−	−	3

Full Factorial Design (2^3) repeated within each batch. Randomized order of the 8 experimental runs within each batch.

Repetitions $\Longrightarrow$ 14 residual degrees-of-freedom.

When doing inference (confidence intervals + tests) the variation between batches is moved from [I] to [batch] so that [I] primarily contains measurement error.

Information in 2^3 design within 3 batches



Fractional Factorial Design

Example: a half-factorial design for 4 dichotomous variables

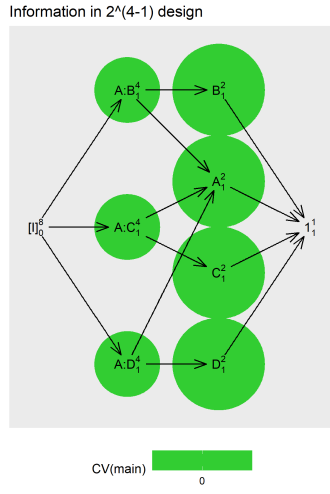
	A	B	C	D
1	1	1	1	1
4	-1	-1	1	1
6	-1	1	-1	1
7	1	-1	-1	1
10	-1	1	1	-1
11	1	-1	1	-1
13	1	1	-1	-1
16	-1	-1	-1	-1

To aid the notation (for the mathematics) the variables are named A, B, C, D, and the levels are named -1 and 1.

From the full factorial design ($2^4 = 16$) we only keep the $2^{4-1} = 8$ experimental runs obeying the **aliasing**

$$A \cdot B = C \cdot D$$

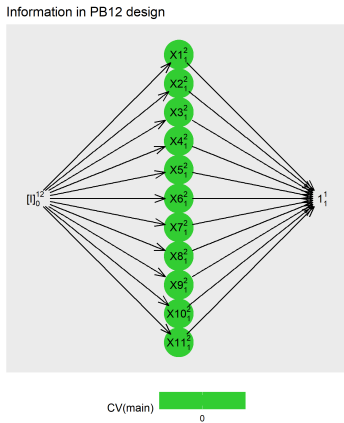
Remark: we can not estimate all interactions.



Plackett-Burman Design (below for $N = 12$)

- **Purpose:** A statistical design allowing for the estimation of $N - 1$ dichotomous (yes/no) variables using N experimental runs.
- **Existence:** If N is a multiple of 4, then the Plackett-Burman design provides a solution to the problem.

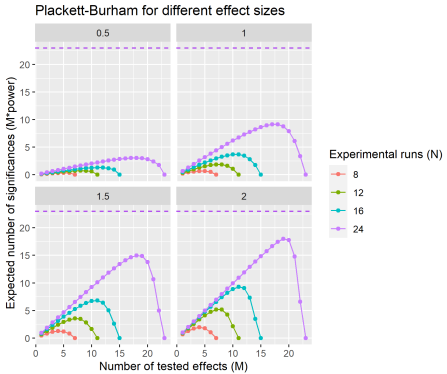
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
1	+	+	−	+	+	+	−	−	−	+	−
2	−	+	+	−	+	+	+	−	−	−	+
3	+	−	+	+	−	+	+	+	−	−	−
4	−	+	−	+	+	−	+	+	+	−	−
5	−	−	+	−	+	+	−	+	+	+	−
6	−	−	−	+	−	+	+	−	+	+	+
7	+	−	−	−	+	−	+	+	−	+	+
8	+	+	−	−	−	+	−	+	+	−	+
9	+	+	+	−	−	−	+	−	+	+	−
10	−	+	+	+	−	−	−	+	−	+	+
11	+	−	+	+	+	−	−	−	+	−	+
12	−	−	−	−	−	−	−	−	−	−	−



Trade-off: more repetitions or more factors?

Assume we are testing M dichotomous factors with difference between the treatments of size δ relative to the residual standard deviation. We do this is a Plackett-Burman design with N experimental runs.

How many significant results can we expect to find?



Quiz: Discuss the trade-off between having more repetitions per factor ($\frac{N}{M}$) and having more factors (M)?

Expected number of findings (used on slide 46)

The expected number of findings is given as $M \cdot \text{power}$, where

$$N = \#\text{ experimental runs}, \qquad M = \#\text{ tested effects}, \qquad \delta = \frac{\text{treatment difference}}{\text{standard deviation}}$$

Assuming standard deviation=1 each comparison in the PB-design has standard error $= \sqrt{\frac{1}{N}}$ with $\text{df} = N - M - 1$ degrees-of-freedom. To align with a t-test (i.e. when one effect is tested), this looks as if the standard deviation equals

$$\tau = \sqrt{\frac{\text{df} + 2}{N}}$$

Thus, given the input (N, M, delta) and the derived (df, tau) the following R code gives the expected number of findings

```
M * power.t.test(n=(df + 2) / 2, delta = delta, sd = tau, sig.level = 0.05)$power
```

Summary

Take home messages

1. In general the residual does not only arise from measurements errors. Instead, the error terms contain all the non-modelled components.
2. Random effects may be used to separate different sources of variation (e.g. animals, batches) out of the residual variation.
3. We had a look at 3 classes of experimental designs: Full factorial, Fractional Factorial, Plackett–Burman.
4. All of these were for dichotomous variables but there are also designs for factors with more than two levels.
5. It is recommended to use orthogonal designs. In general, these are more powerful and resulting estimates are more easily interpretable.
6. Design Diagrams provide visualization of proposed designs.

End of lectures

This concludes the lecture part of the course

- There is a final exercise session this afternoon.
- AS participants do an applied project, and SmS-participants may sign up for an applied project in block 3 (see "Data Science Projects" on `https://datalab.science.ku.dk/english/course/`):
  - Using statistical methodology in practice is the best way to really learn statistics.
  - Doing statistics in practice may be surprisingly difficult. It involves biological and mathematical reasoning, as well as technical skills. Its is wonderfully challenging! You may keep on learning for the rest of your lives ☺.

Thank you!