Applied Statistic

Chapter 8 - Checking the models I : independence

Christian Damgaard
Ecoscience
Aarhus University

Assumptions of GLM

All statistical tests rely on assumptions!

GLM are parametric tests with the following four assumptions:

- Independence
- Homogeneity of residual variance
- Normality of residual variance
- Linearity / additivity

What happens if assumptions are violated?

- Incorrect parameter estimates
  - biased estimates of mean and variance
- Incorrect conclusion on statistical significance
  - wrong p-value

Assumption of independence

Given a random sample (y_i, x_i) for $i = 1,2, \dots, n$

and a liner model, $y_i = \mu + \alpha\, x_i + \epsilon_i$

Then we may define independence as: $P(\epsilon_i) = P(\epsilon_i|\epsilon_j)$

Datapoints are independent in a given linear model if knowing the residual error of one or a subset of datapoints provides no knowledge of the residual error of any other

Assumption of independence

Independence is a key assumption, and the most problematic in practice

Lack of independence is the single most important cause of serious statistical problems

Realize that mistakes at the design stage are often unrecoverable at analysis – think before sampling!

The experimental unit

Ensure random sampling at the level of the experimental unit

Pseudo-replication: number of measurements are higher than the number of experimental units

Problems with independence in hierarchical data structures

Repeated data: the same experimental unit is measured more times
the weight of 5 pigs has each been measured 4 times
what is the sample size?

Nested data: groups of data
10 sites with each 10 plots
what is the sample size?
what is the sample size if we do not detect an effect of site?

Sheep example (Exercise 8-1)

The lookup rate was observed 20 times for each of 3 male sheep and 3 female sheep

Did SEX have an affect on the lookup rate?

```
> lm1 <- lm(LUPRATE ~ SEX + SHEEP, data=sheepdata)
> Anova(lm1, type="II")
Anova Table (Type II tests)

Response: LUPRATE
           Sum Sq  Df F value    Pr(>F)
SEX        0.36130   1 131.849 < 2.2e-16 ***
SHEEP      0.29477   4  26.893 1.015e-15 ***
Residuals 0.31239 114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.
```

LUPRATE=SEX+ SHEEP

But SEX is nested within SHEEP, so

LUPRATE=SEX+ SHEEP is a **wrong** model

SHEEP has 4 Df instead of 5 Df

114 residuals Df !!!

Instead use mean values or make a nested model (chapter 12)

The experimental unit and nested design

Sow rate	1	2	3	4
Variety 1	3	3	3	3
Variety 2	3	3	3	3

In the bean yield example it was possible to investigate all combinations of sow rate and variety

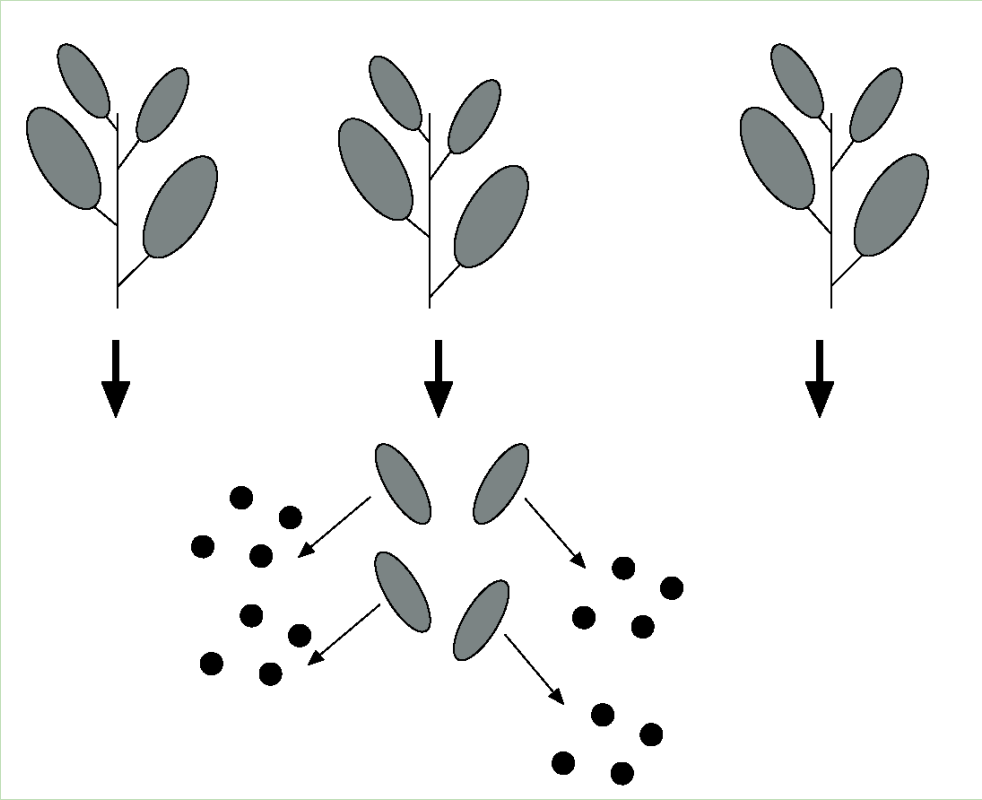
Sheep	1	2	3	4	5	6
Female	20	20	20			
Male				20	20	20

However, when analyzing the effect of sex on sheep behavior it is not possible to randomize sex for a sheep

Sheep is the experimental unit (nested design)

Nested design: calcium content in leaves example

Three plants * four leaves* four discs = 48 measurements



Why use a nested design?

What is the experimental unit?
What is the sample size?

How to analyze a nested design ?

- Use the mean values
- A mixed-effect model (chap.12)

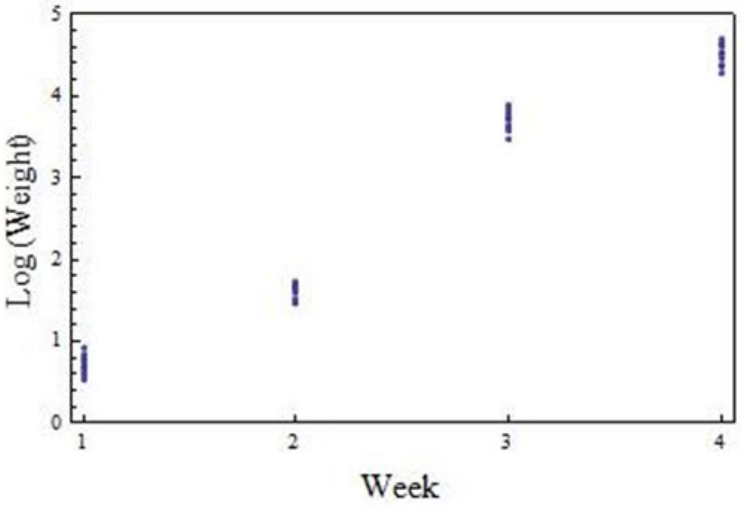
Repeated measures: pig example

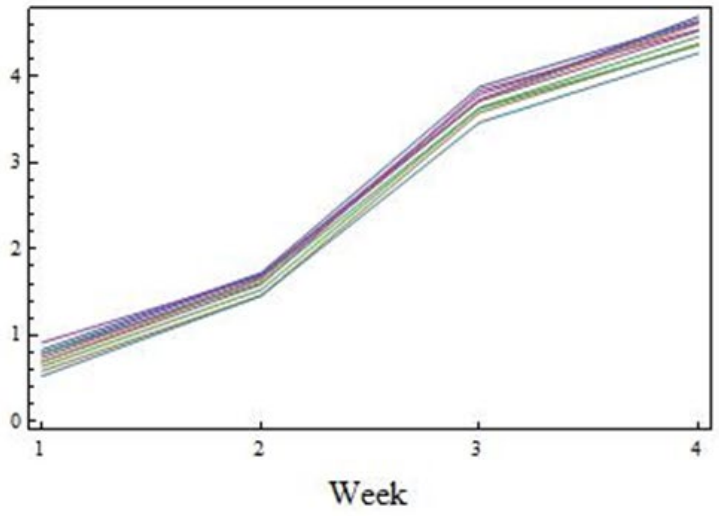
DIET	PIG	SAMPLE	LGWT	DIET	PIG	SAMPLE	LGWT
1	1	1	0.78846	2	6	1	0.74194
1	1	2	1.70475	2	6	2	1.66771
1	1	3	3.72810	2	6	3	3.71357
1	1	4	4.68767	2	6	4	4.52504
1	2	1	0.69315	2	7	1	0.58779
1	2	2	1.58924	2	7	2	1.45862
1	2	3	3.83298	2	7	3	3.58074
1	2	4	4.53903	2	7	4	4.37450
1	3	1	0.69315	2	8	1	0.64185
1	3	2	1.64866	2	8	2	1.52606
1	3	3	3.73050	2	8	3	3.62700
1	3	4	4.60517	2	8	4	4.35927
1	4	1	0.78846	2	9	1	0.53063
1	4	2	1.60944	2	9	2	1.45862
1	4	3	3.63495	2	9	3	3.46574
1	4	4	4.45783	2	9	4	4.26690
1	5	1	0.83291	2	10	1	0.91629
1	5	2	1.72277	2	10	2	1.68640
1	5	3	3.87743	2	10	3	3.79098
1	5	4	4.64150	2	10	4	4.62301

10 pigs were fed (5 with each diet)

Weight gain was measured after 3, 8, 20, 60 weeks

Pig example - what is the experimental unit?



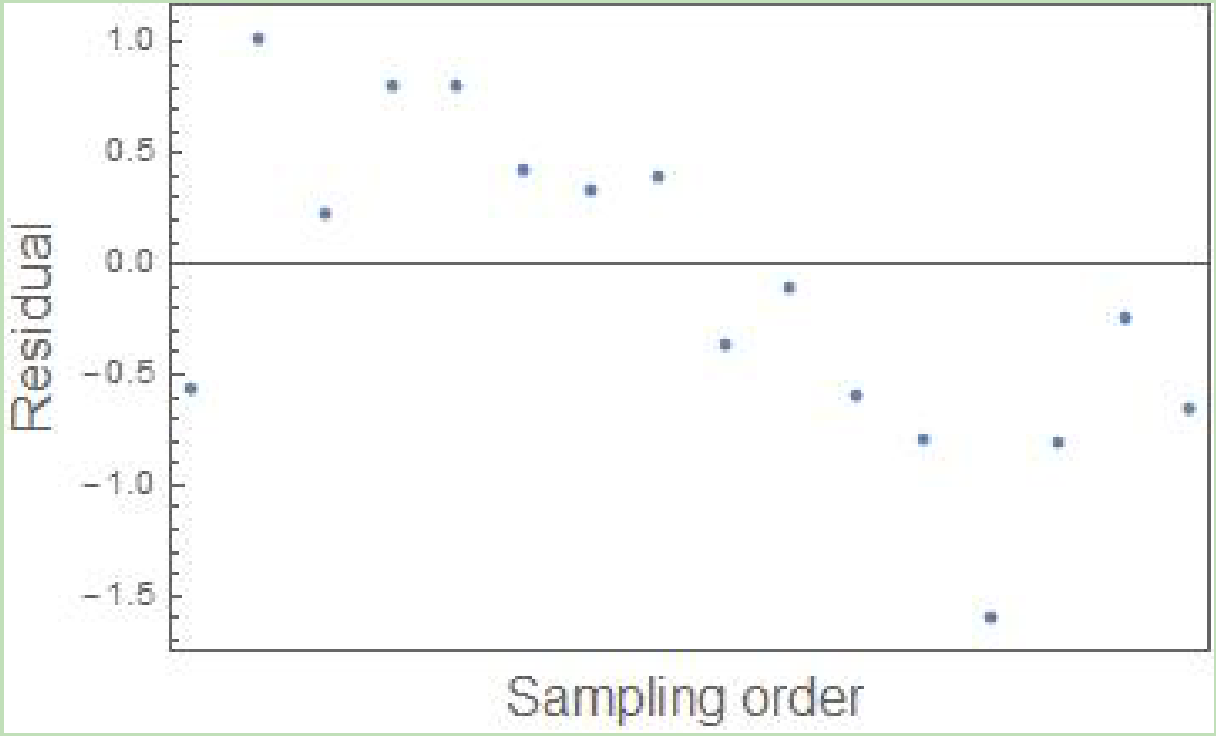


The residuals made from the data in the left panel are not independent – a pig that is relatively large in week 1 tend to be relatively large in the following weeks as well

Time series data

Independent residuals?

Plot residuals as a function of their sampling order



How to analyze time series data?

- Single summary approach - one-pig-one data point
- Repeated measure analysis - a mixed-effect model (chapter 12)
- Multivariate approach (another course)
- Dynamic models or growth models (another course)

Design and independence

15 minutes discussion with neighbor

Realize that mistakes at the design stage are often unrecoverable at analysis – think before sampling!

Design three experiments to measure the effect of adding nitrogen to a natural habitat on plant communities (make notes)

1. Independent and randomized
2. Nested design
3. Repeated measures

Discuss possible drawbacks and advantages of the three designs

Discuss sample sizes – what is realistic in a master project

Design and independence

What measurements?

A single (dominant) species or all higher plants?

Plant abundance or occurrence?

1. Independent and randomized
   a) simple design and statistical analysis
   b) relatively high statistical power with a fixed number of observations
   c) can the results be generalized?
2. Nested design
   a) more habitats / populations – higher degree of generality
   b) hierarchical statistical analysis
3. Repeated measures
   a) needed for measuring growth
   b) non-destructive sampling method is needed

More thoughts?

Important points

Pseudo-replication may lead to huge mistakes