

SS7-I The statistical analysis of multi-site data (part 1-I)

Introduction and motivation

Name: Rouven Lattner
Institute: Trinity College Dublin
Supervisor: Prof. Caroline Brophy

Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

28.05.2026, Wageningen, The Netherlands

Funded by
the European Union

Racines d'avenir
LABORATOIRE VIVANT

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
State Secretariat for Education,
Research and Innovation SERI

UKRI
UK Research and Innovation

Part 1 – Introduction and Motivation

Examples: Grouped data

What motivates the usage of mixed-effects models?

- Observations not just a random collection of independent points
- They can come in groups
- Observations from the same group can share certain characteristics

Observation	Grouping level	Similarity
Students	Schools	Same teacher, school environment, curriculum
Patients	Doctors/Hospital	Same treatment, protocol, same local population
Patients (repeated measures)	Doctors/Hospital	Multiple observations per patient
Field plots	Site/Block	Same soil, weather, management
Animals	Farm/Herd	Management, food, environment

Why independence matters

A standard linear model assumes independent errors

- What are the model assumptions?
  - Linear relationship
  - Normally distributed errors
  - Constant error variance
  - **Independence**

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

$$\rightarrow Cov(\epsilon_i, \epsilon_j) = 0 \quad for\ i \neq j$$

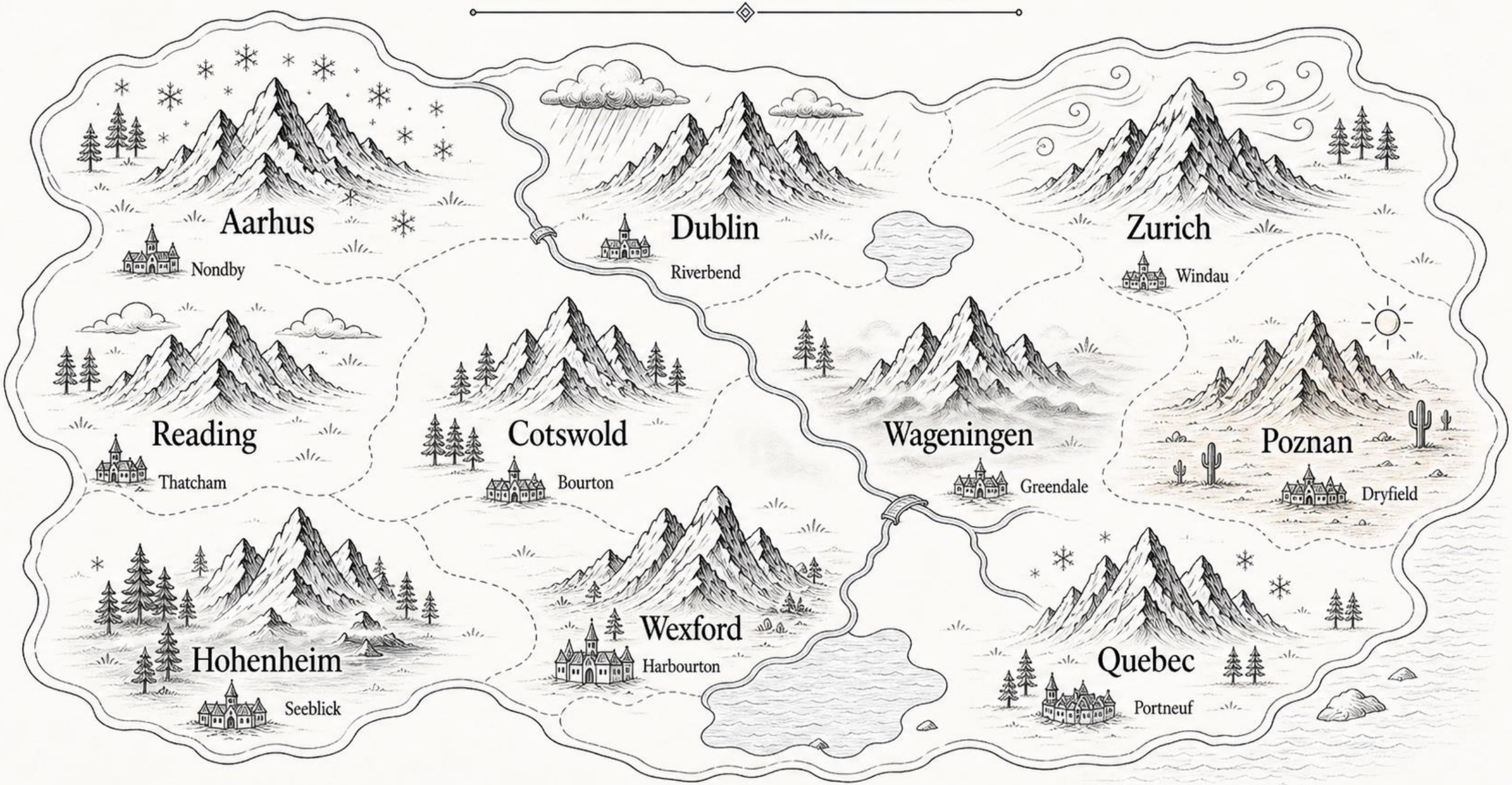
- **Observations within the same group can be more similar than expected**
- **Residuals may be correlated**
- **Pseudo replication: 250 dragons from 10 ranges ≠ 250 independent replicates**

Tox example: Dragon transport data

Background context for simulated data

- Before the invention of machines, people utilised animals to assist with farming
- In a mountainous farming world, dragons were used to transport harvested crop material to nearby cities
- We record how much crop mass each dragon carries per day
- We can also record dragon characteristics (weight, wing span, age ...)
- What are characteristic that influence the performance of dragons?
- How do they differ across mountain ranges?

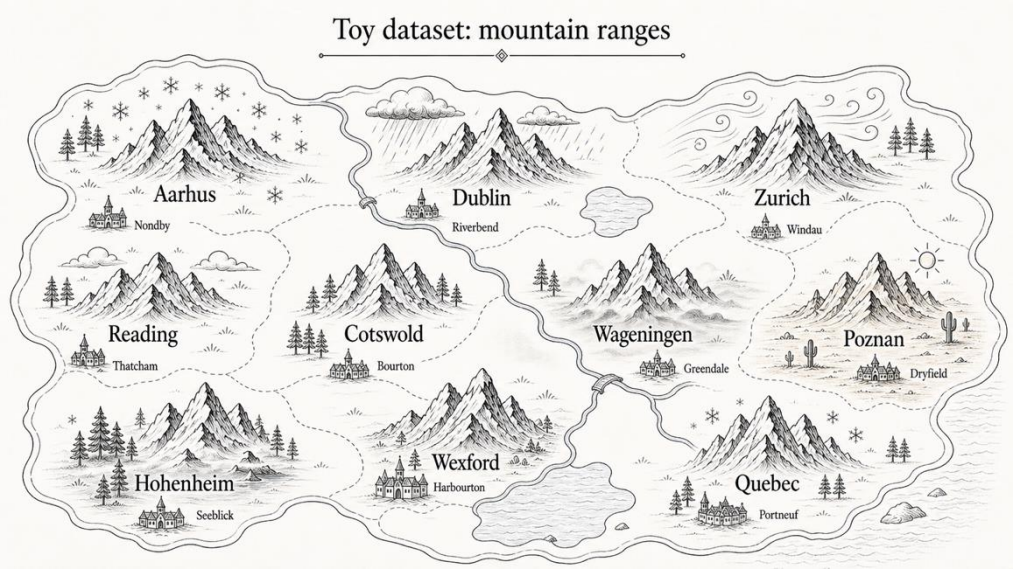
Toy dataset: mountain ranges



Tox example: Dragon transport data

Structure

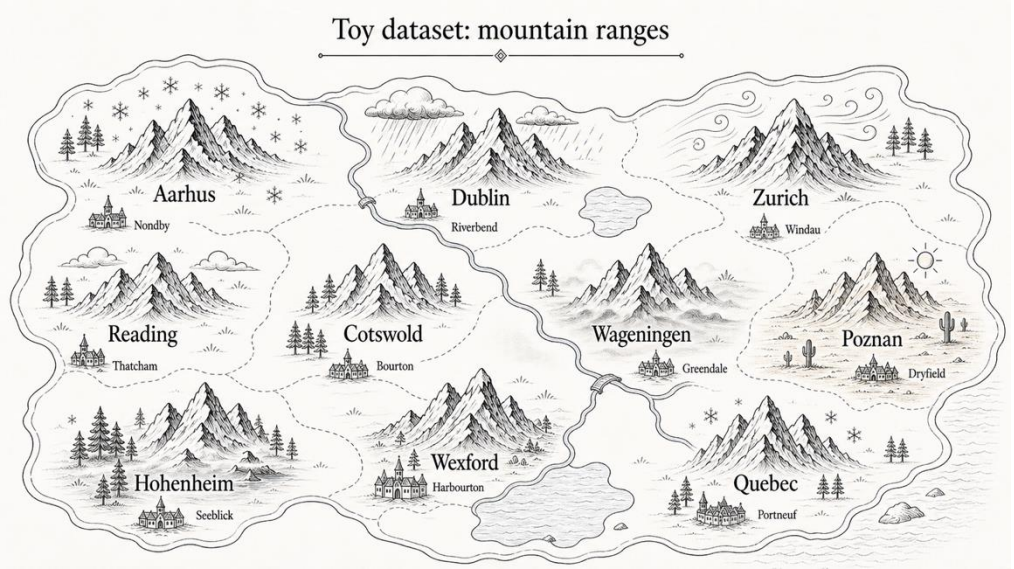
- One row = one unique dragon
- Response: Carried crop mass per day
- Predictor: Body weight
- Grouping variable: Mountain range
- 25 observations per range → 250 observations



Dragon ID	Mountain range	Weight (kg)	Carried mass (kg)
1	Dublin	642	585
2	Zurich	781	755
3	Aarhus	524	468
…	…	…	…

Tox example: Dragon transport data

Structure

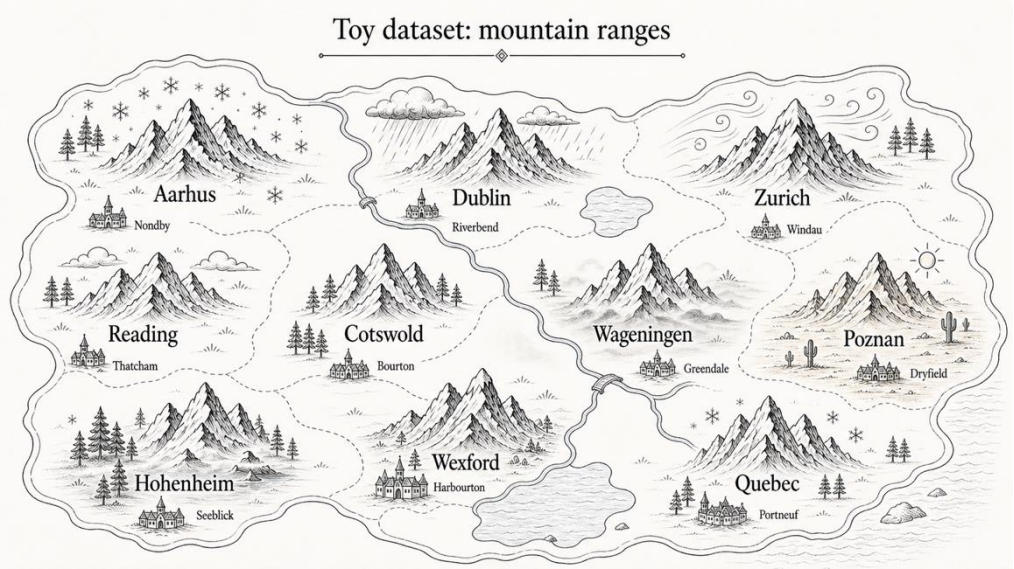


- One row = one unique dragon
- Response: Carried crop mass per day
- Predictor: Body weight
- Grouping variable: Mountain range
- **Mountain ranges in our data are treated as a sample from a wider population of ranges**

- **Do heavier dragons carry more crop mass per day?**

Tox example: Dragon transport data

Structure

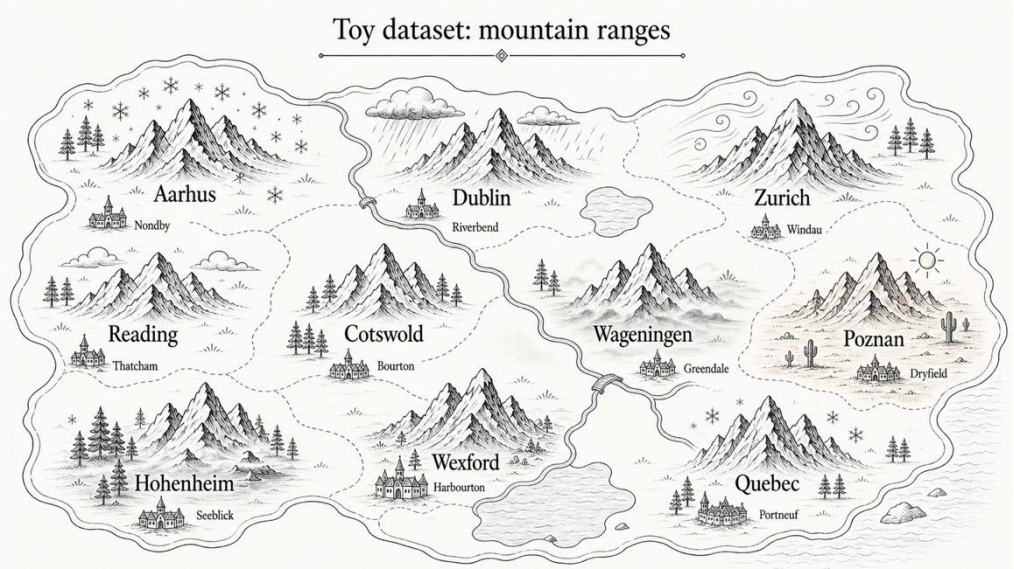


- One row = one unique dragon
- Response: Carried crop mass per day
- Predictor: Body weight
- Grouping variable: Mountain range
- **Mountain ranges in our data are treated as a sample from a wider population of ranges**

- **Do heavier dragons carry more crop mass per day?**
  - What is the average performance? (Population average)
  - Are there performance differences between sites?
  - If a new mountain range appears, how would dragons from this range perform?

Tox example: Dragon transport data

Structure



- One row = one unique dragon
- Response: Carried crop mass per day
- Predictor: Body weight
- Grouping variable: Mountain range
- **Mountain ranges in our data are treated as a sample from a wider population of ranges**

- **Do heavier dragons carry more crop mass per day?**
  - What is the average performance? (Population average)
  - Are there performance differences between sites?
  - If a new mountain range appears, how would dragons from this range perform?
- **Can we answer that with a regular linear regression model?**

Tox example: Dragon transport data

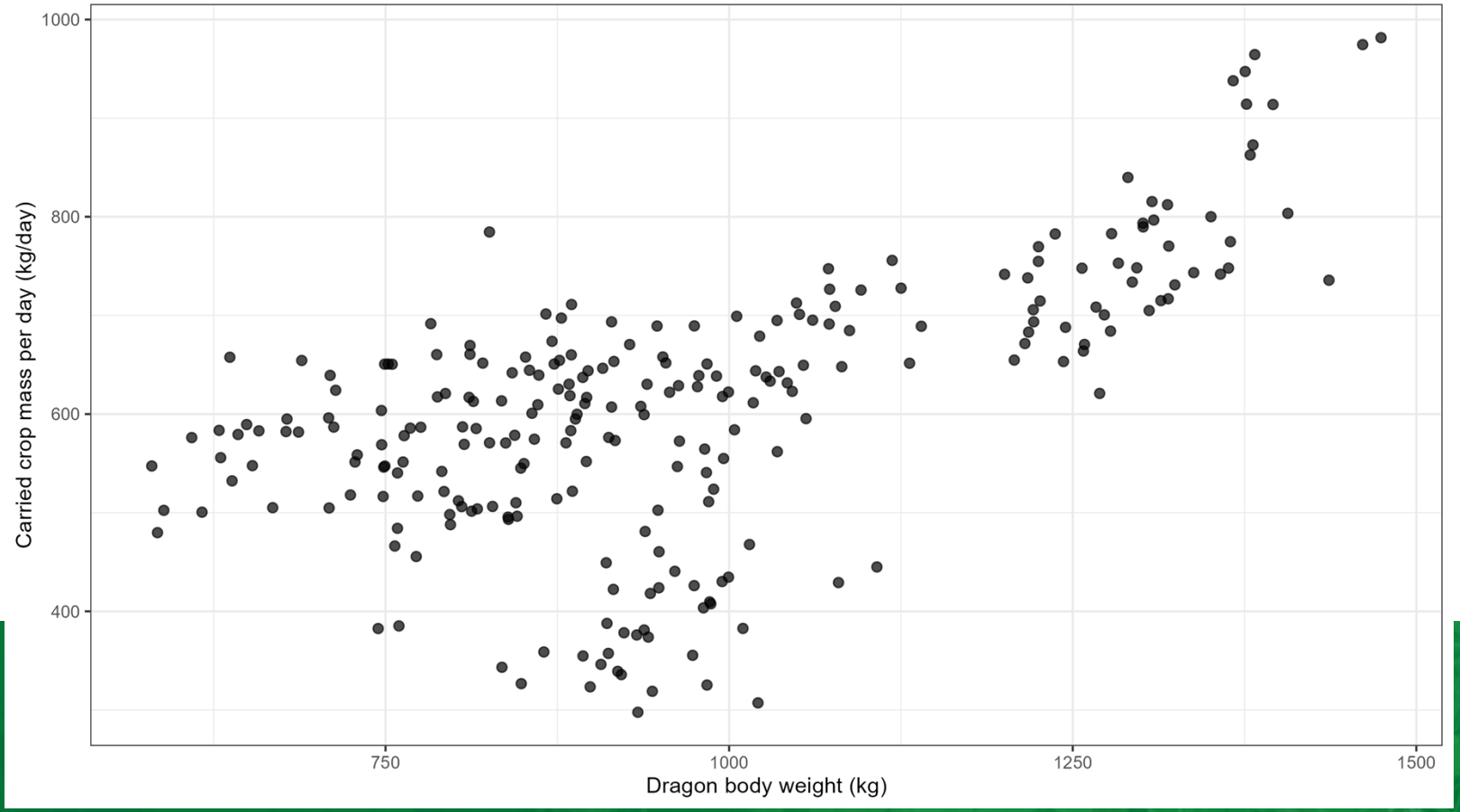
Exploring the data

- Variety of different plots to answer this question
- Exploratory analysis can reveal the need for mixed-effects models

Tox example: Dragon transport data

Exploring the data – First look

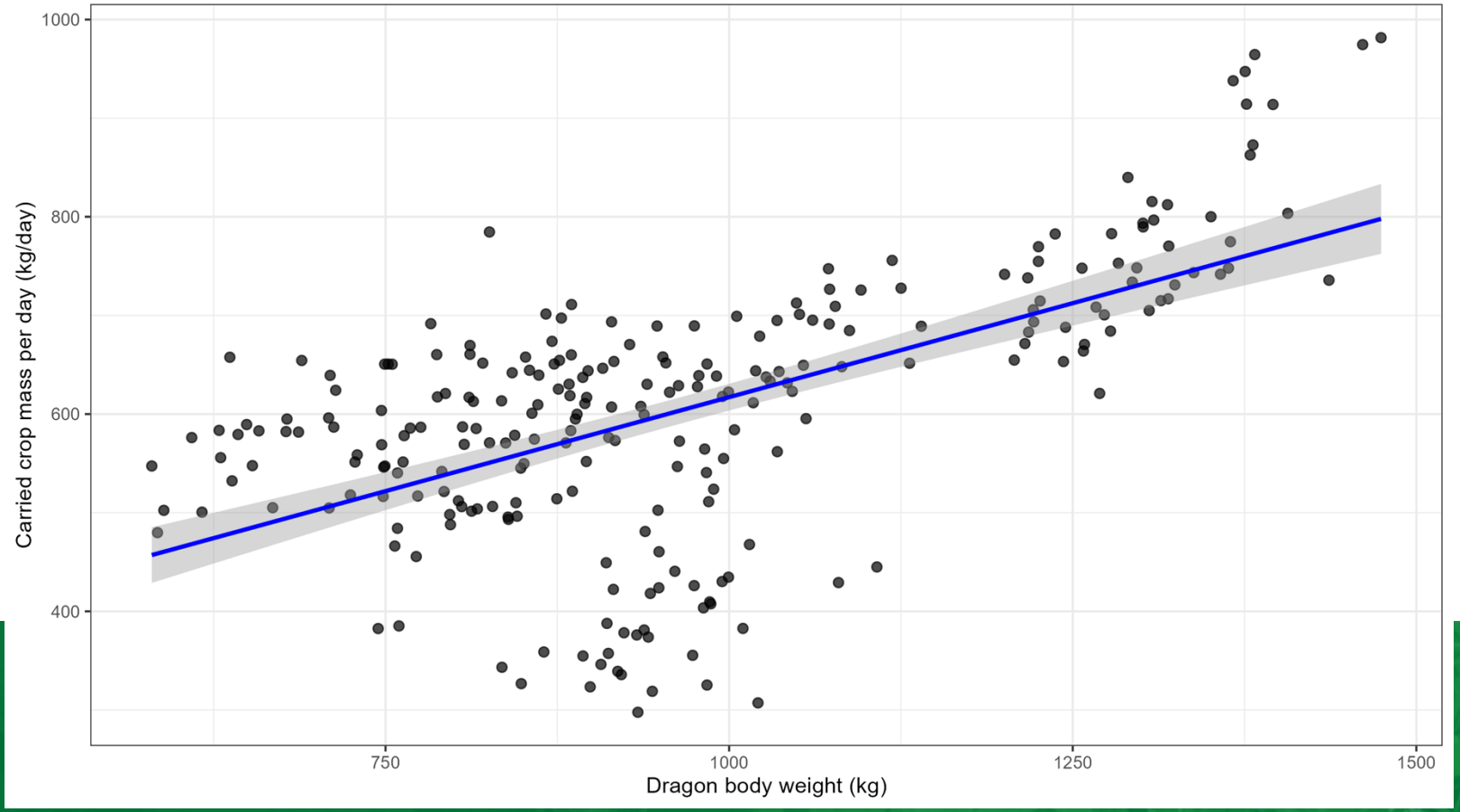
- **Do heavier dragons carry more crop mass per day?**



Tox example: Dragon transport data

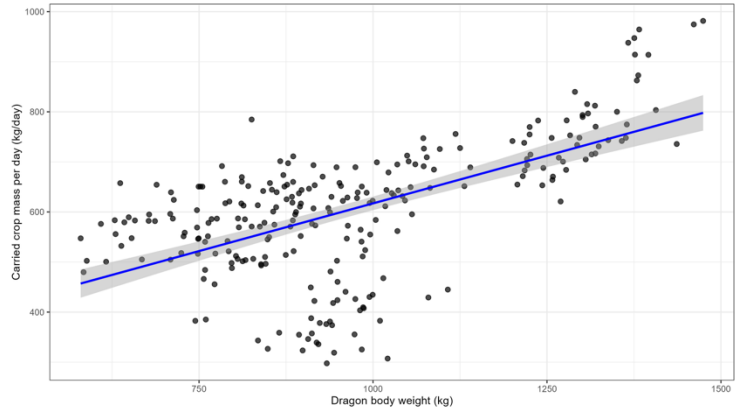
Exploring the data – Simple linear regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

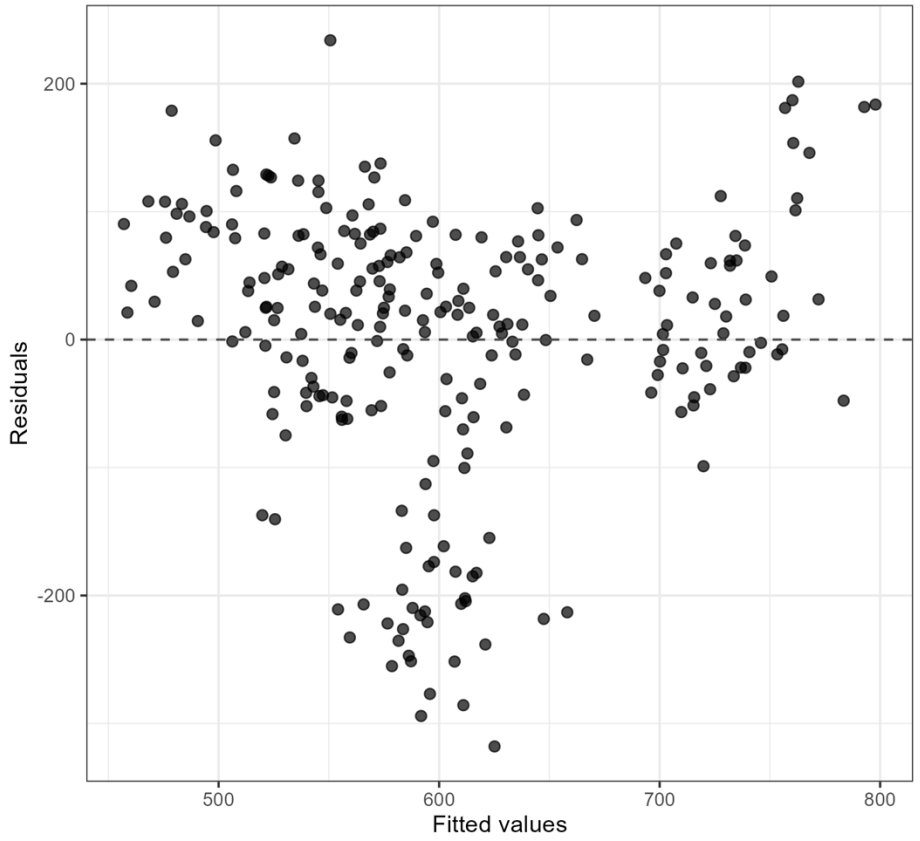


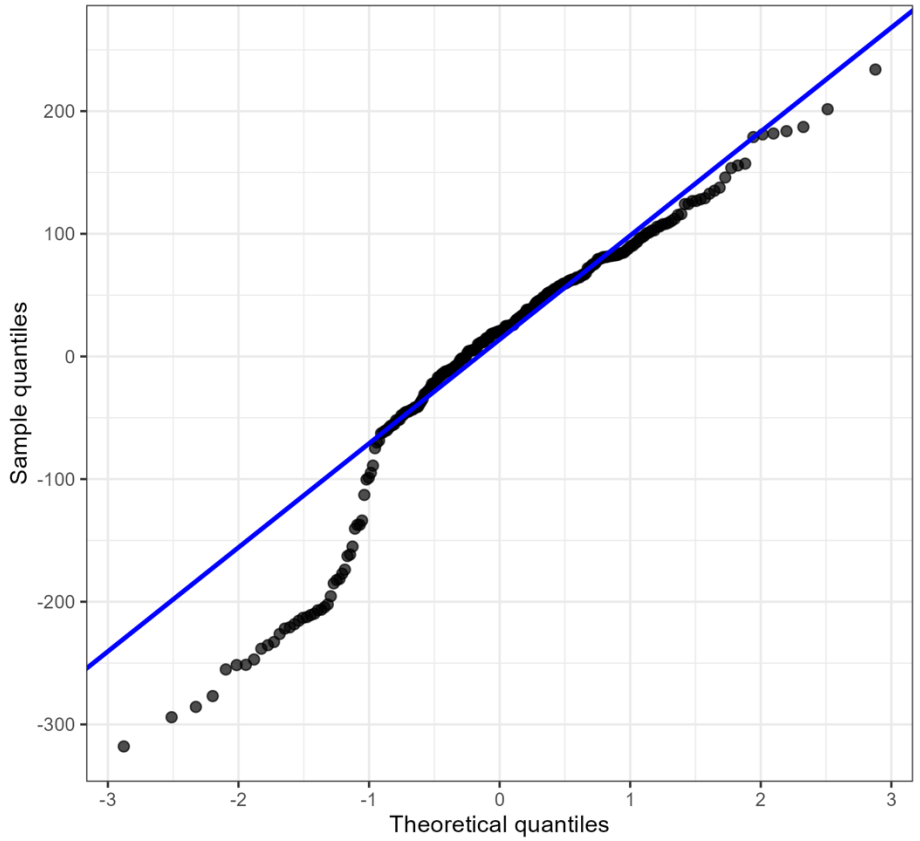
Tox example: Dragon transport data

Exploring the data – Simple linear regression assumptions



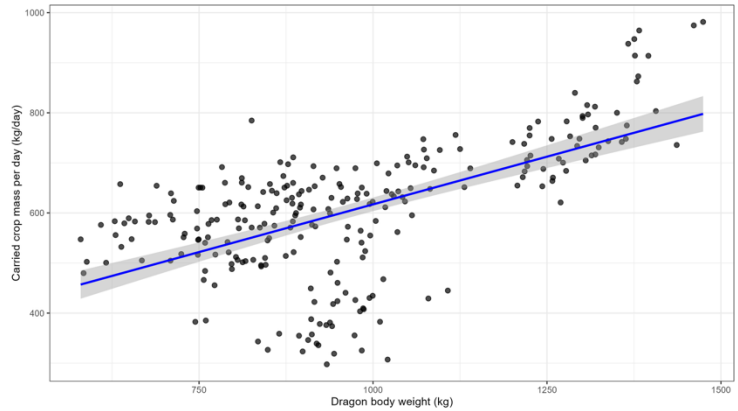
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$



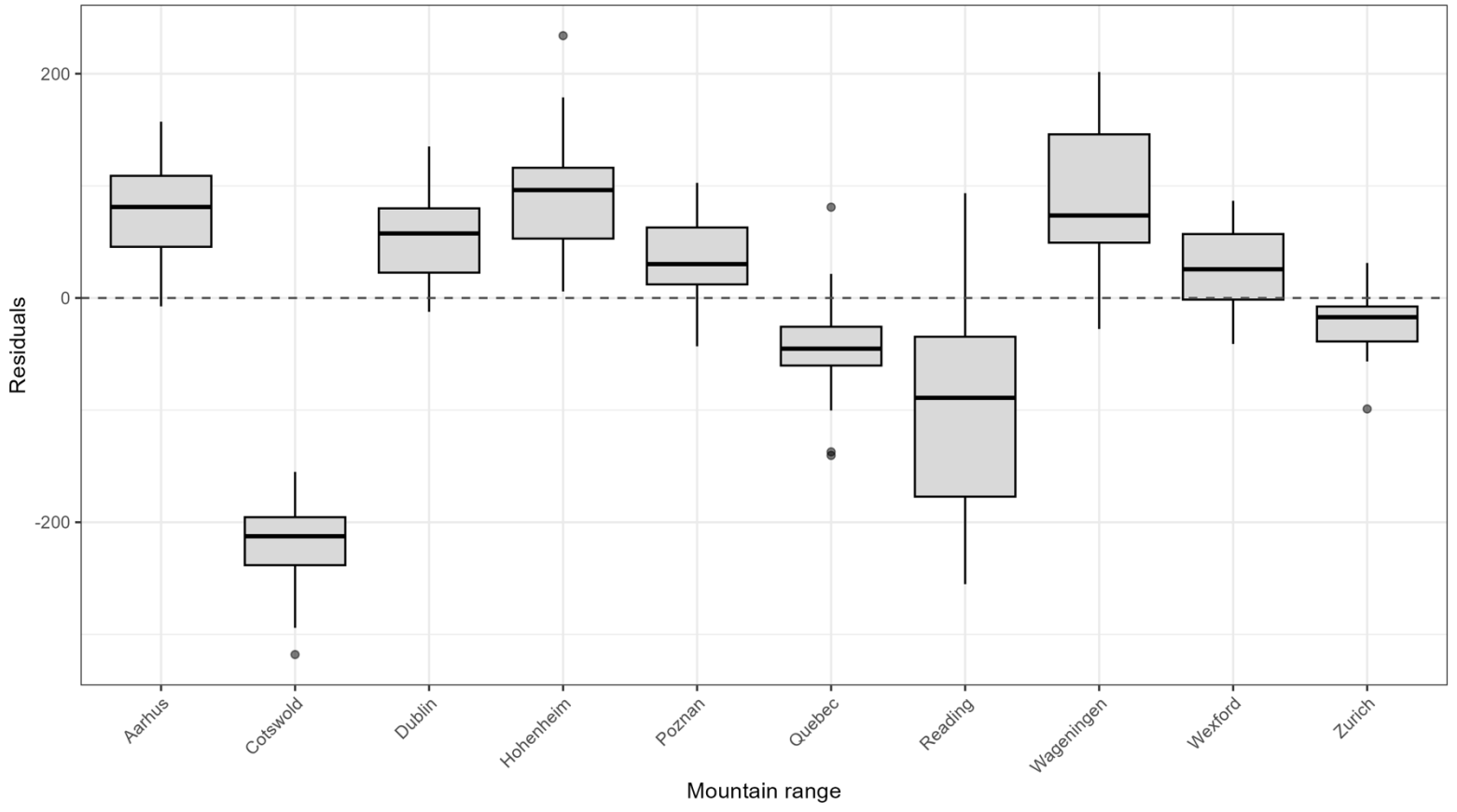


Tox example: Dragon transport data

Exploring the data – Simple linear regression group residuals



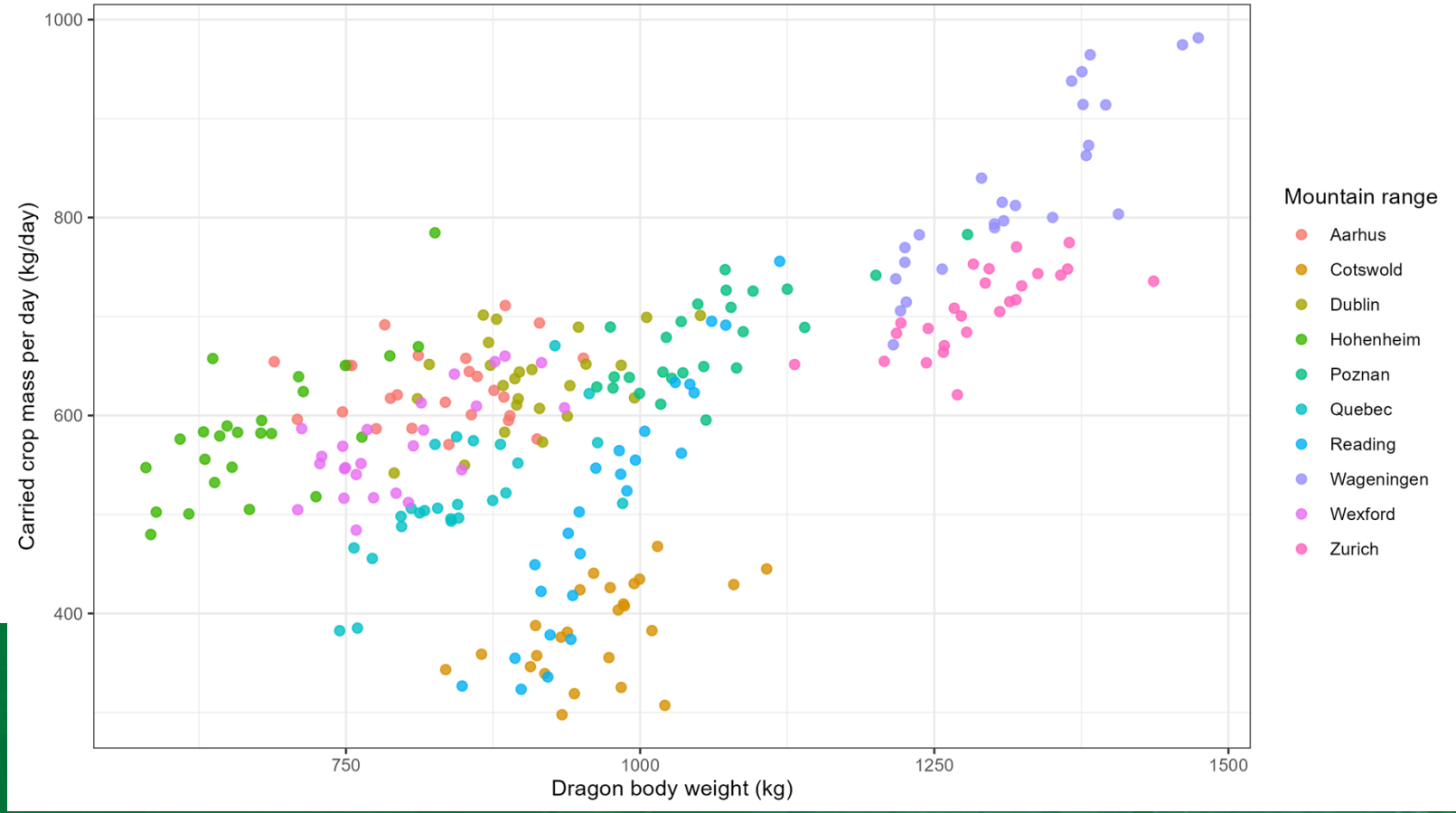
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$



Tox example: Dragon transport data

Exploring the data – Adding colours for grouping structure

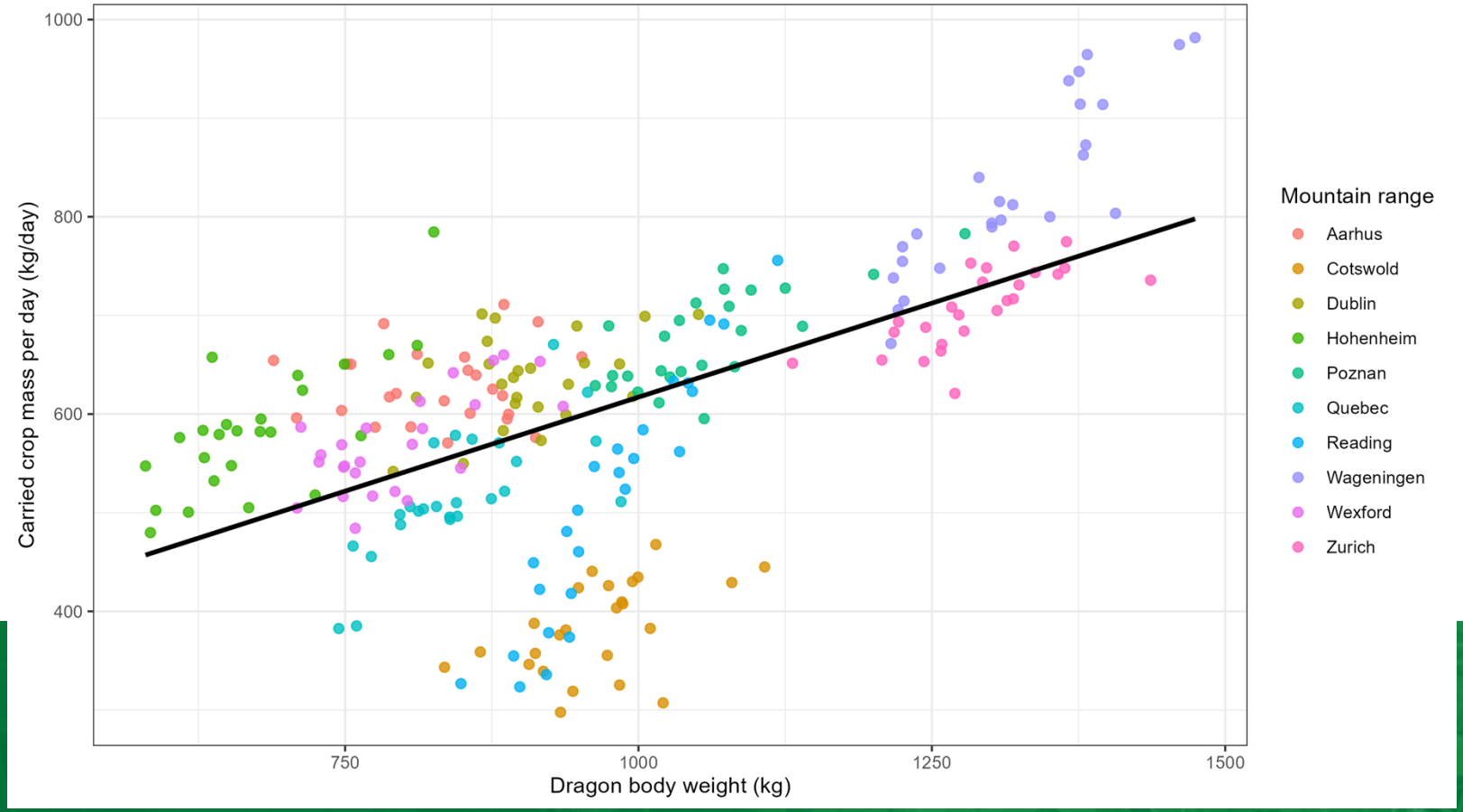
- **Do heavier dragons carry more crop mass per day?**



Tox example: Dragon transport data

Exploring the data – Adding regression line to grouping structure

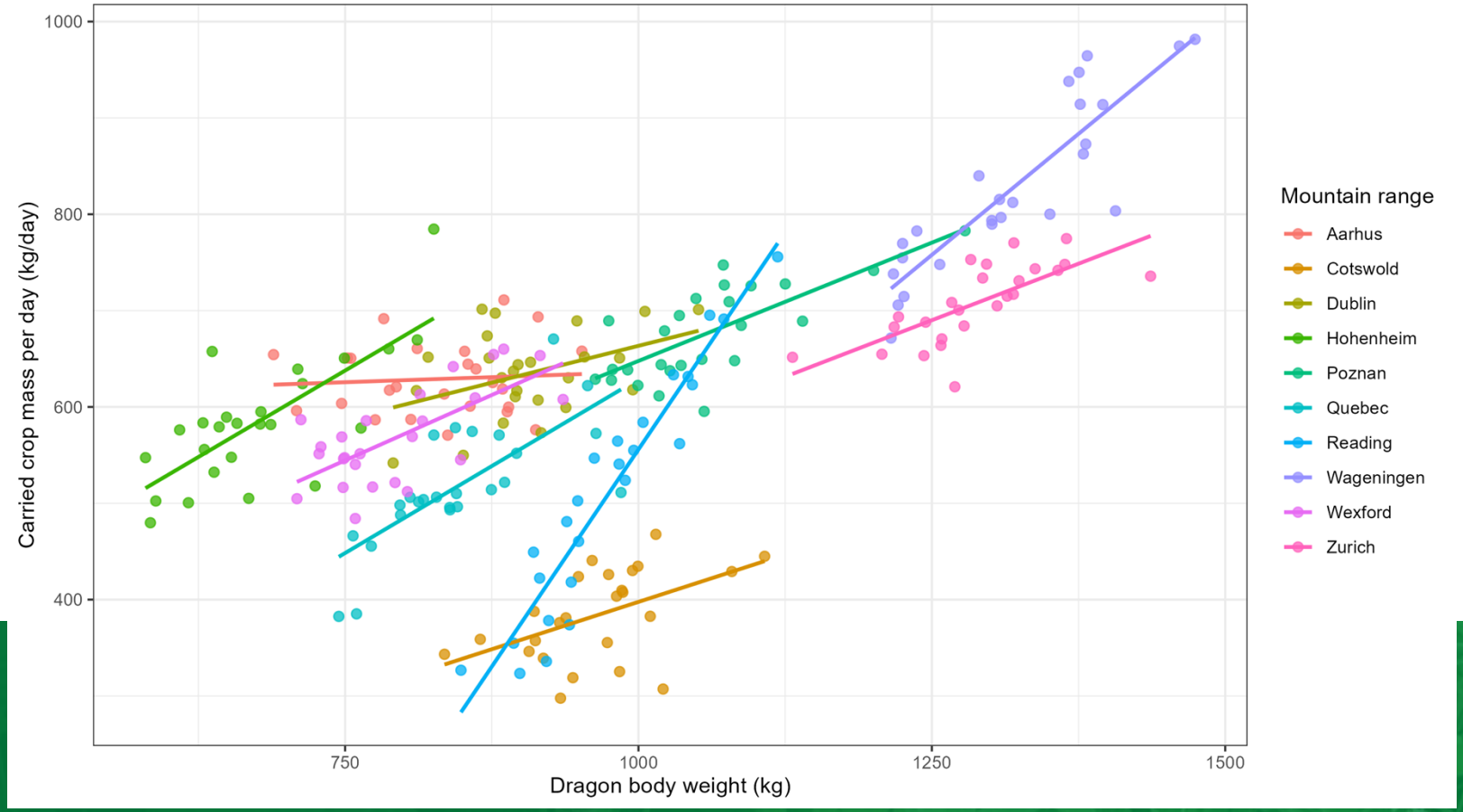
- **Do heavier dragons carry more crop mass per day?**



Tox example: Dragon transport data

Exploring the data – Individual regression line per mountain range

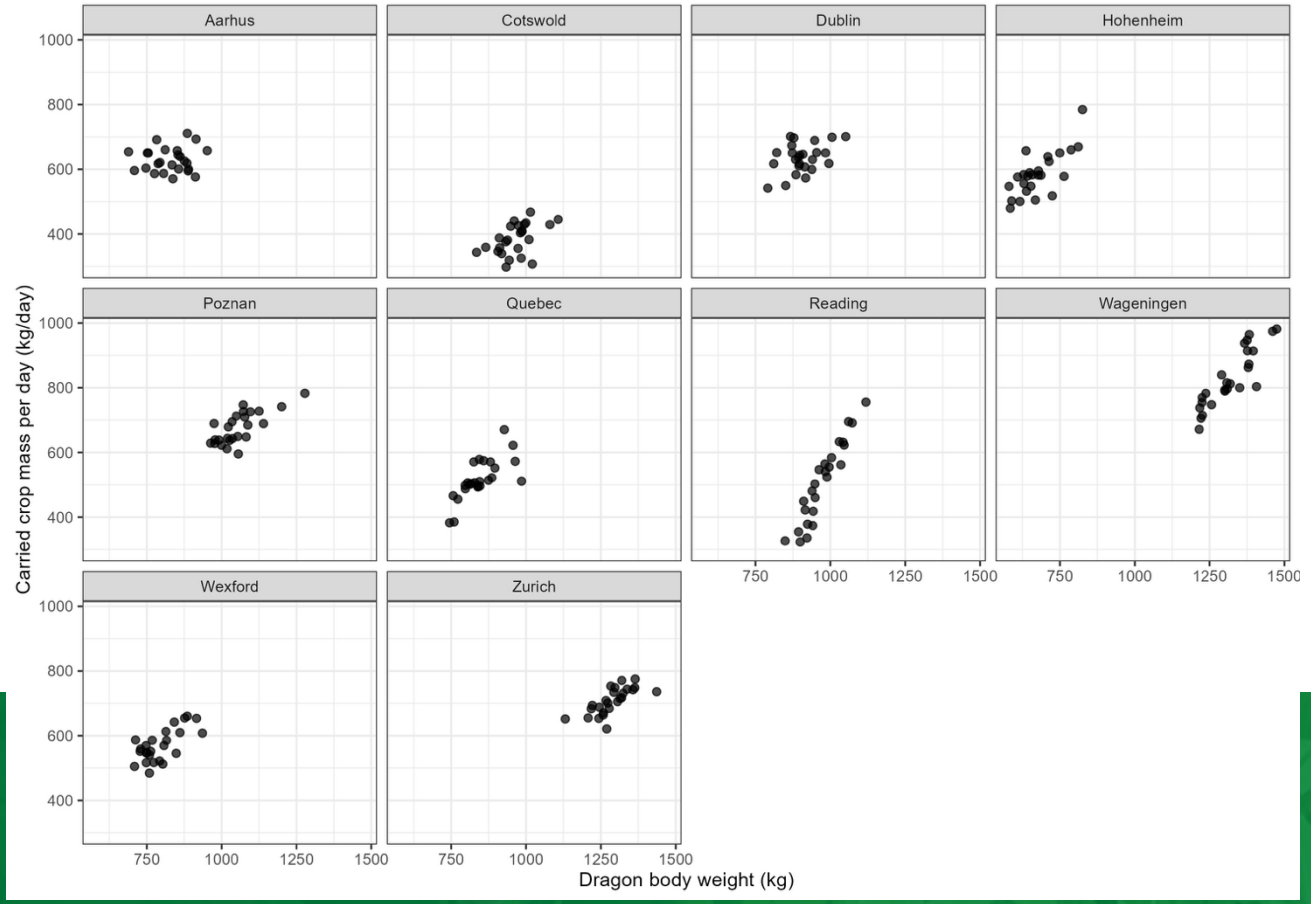
- **Do heavier dragons carry more crop mass per day?**



Tox example: Dragon transport data

Exploring the data – Individual plot per mountain range

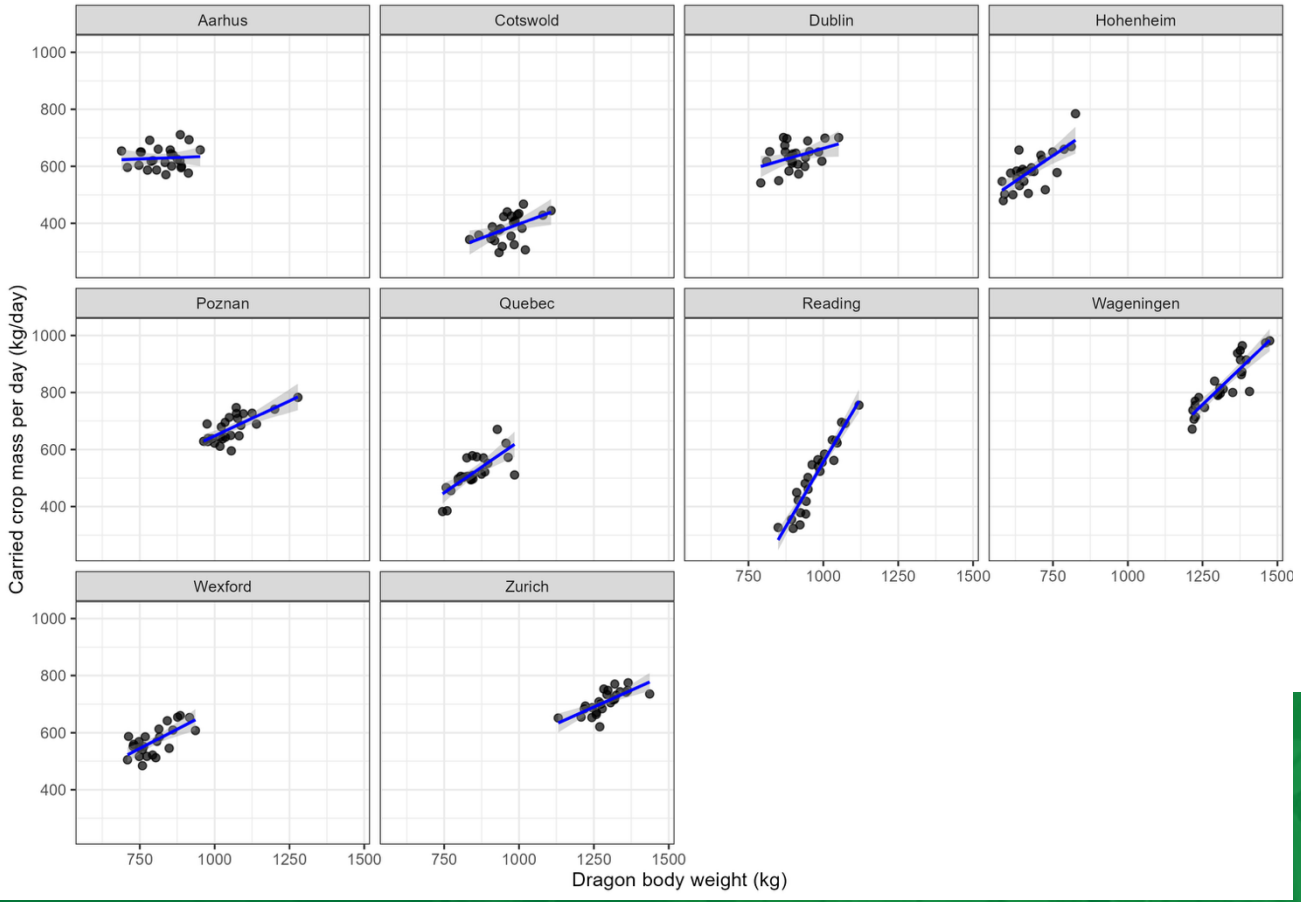
- **Looking into observations from each mountain range**



Tox example: Dragon transport data

Exploring the data – Individual plot per mountain range

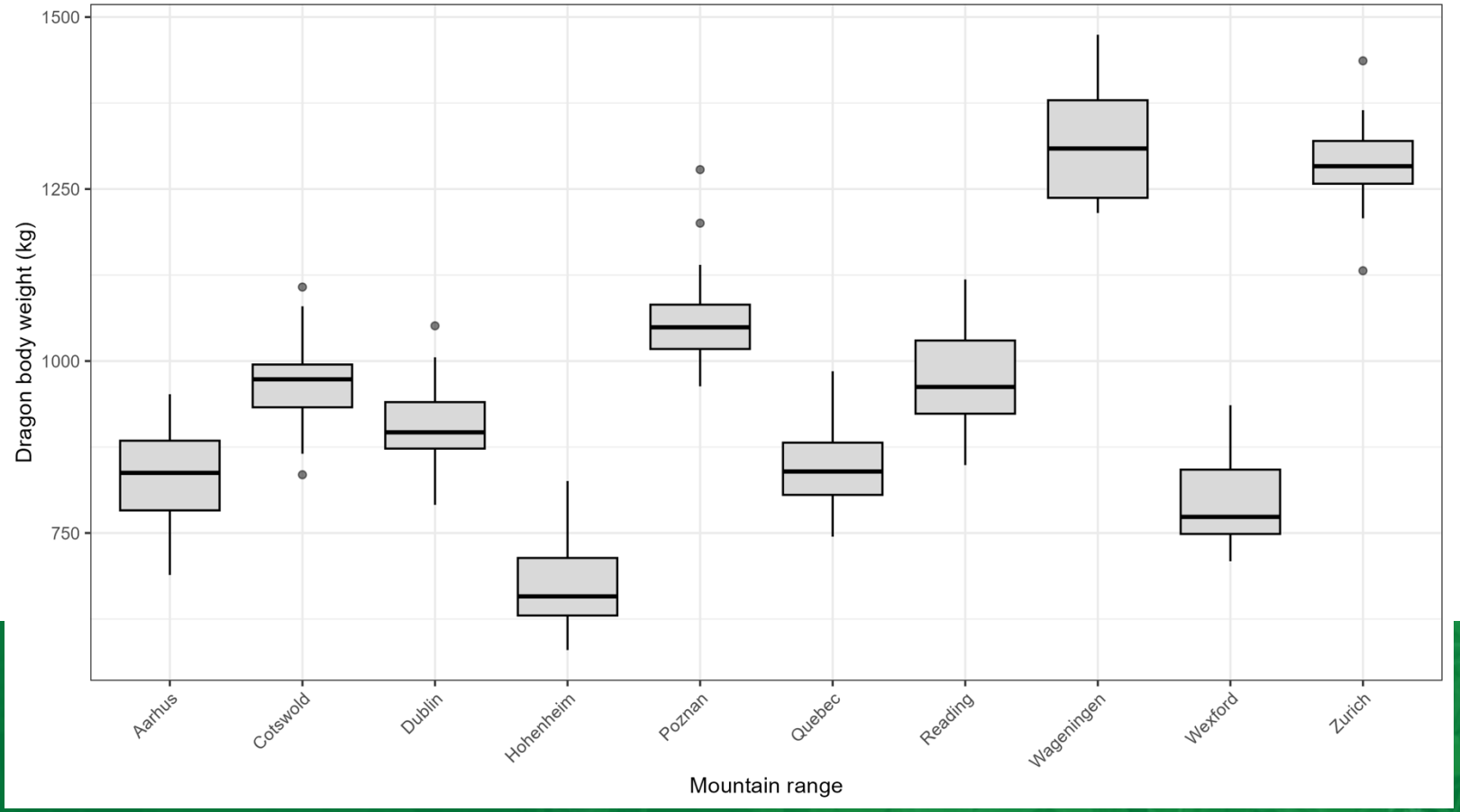
- **Looking into observations from each mountain range**
- **Adding regression lines**



Tox example: Dragon transport data

Exploring the data – Differences by group

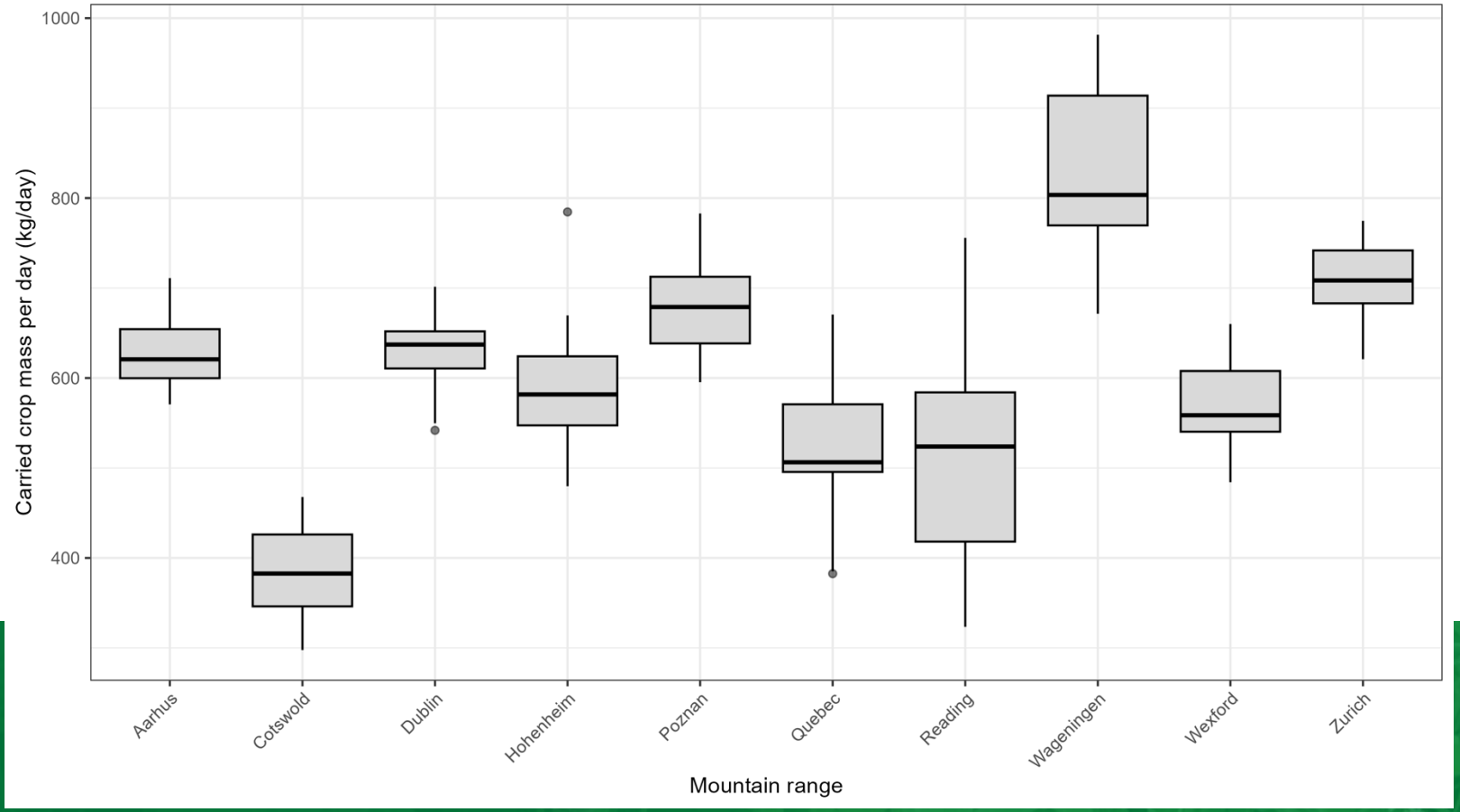
- **Mountain ranges differ in dragon body weight**



Tox example: Dragon transport data

Exploring the data – Differences by group

- **Mountain ranges differ in carrying capacity**



Modelling strategies

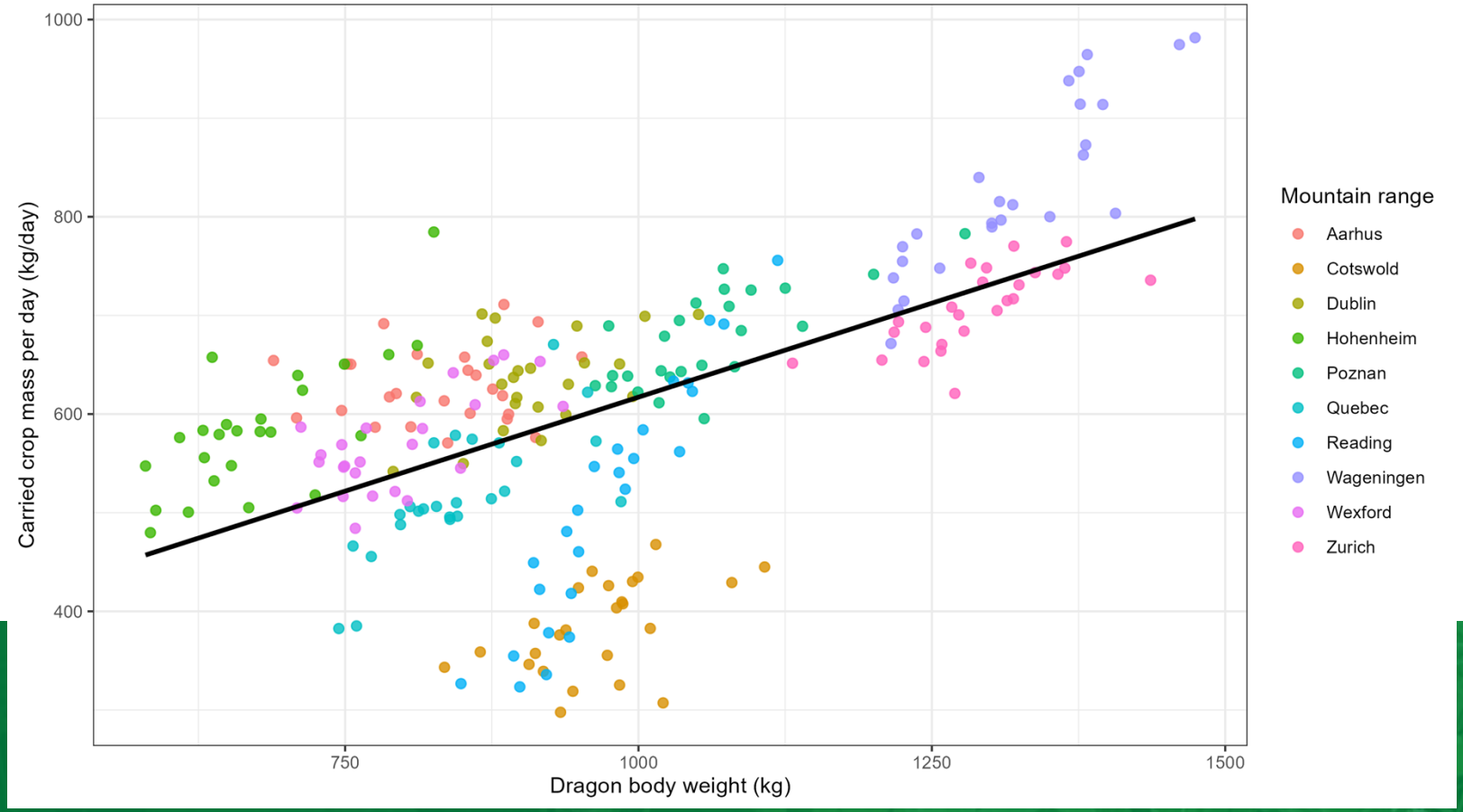
Overview of possibilities

- **What could we do instead?**
  - Ignore groups
  - Aggregate data
  - Fit individual models
  - Add “mountain range” as a factor variable
  - Linear mixed-effects models

Modelling strategies

Ignore groups

- **Simple linear regression**



Modelling strategies

Aggregate data

- Many dragons per range → One average observation per range

Can be reasonable if question is about range-level averages, BUT:

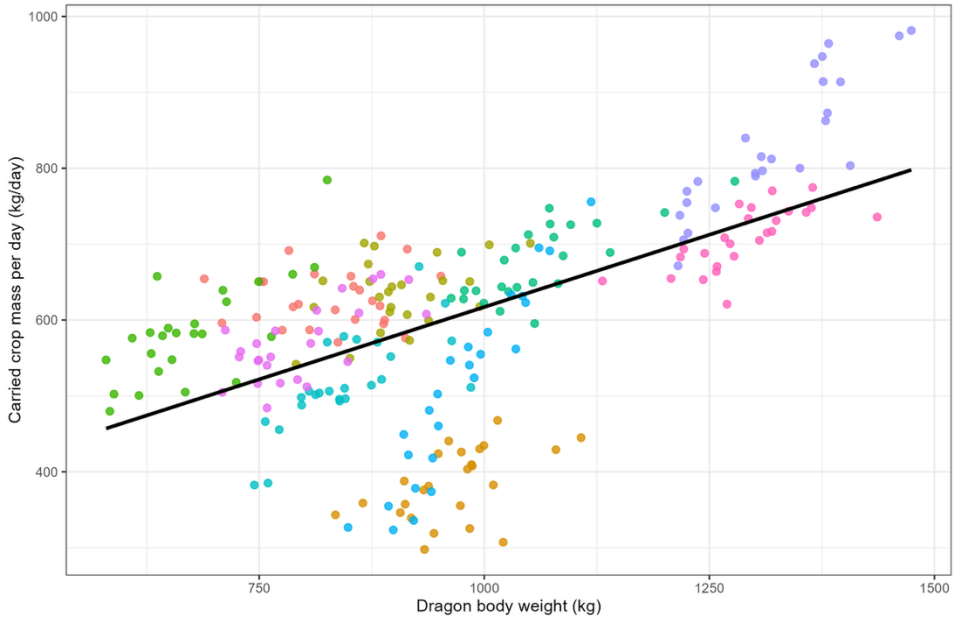
- The research question changes
- Within-range relationship can no longer be studied
- Uncertainty now based number of mountain ranges

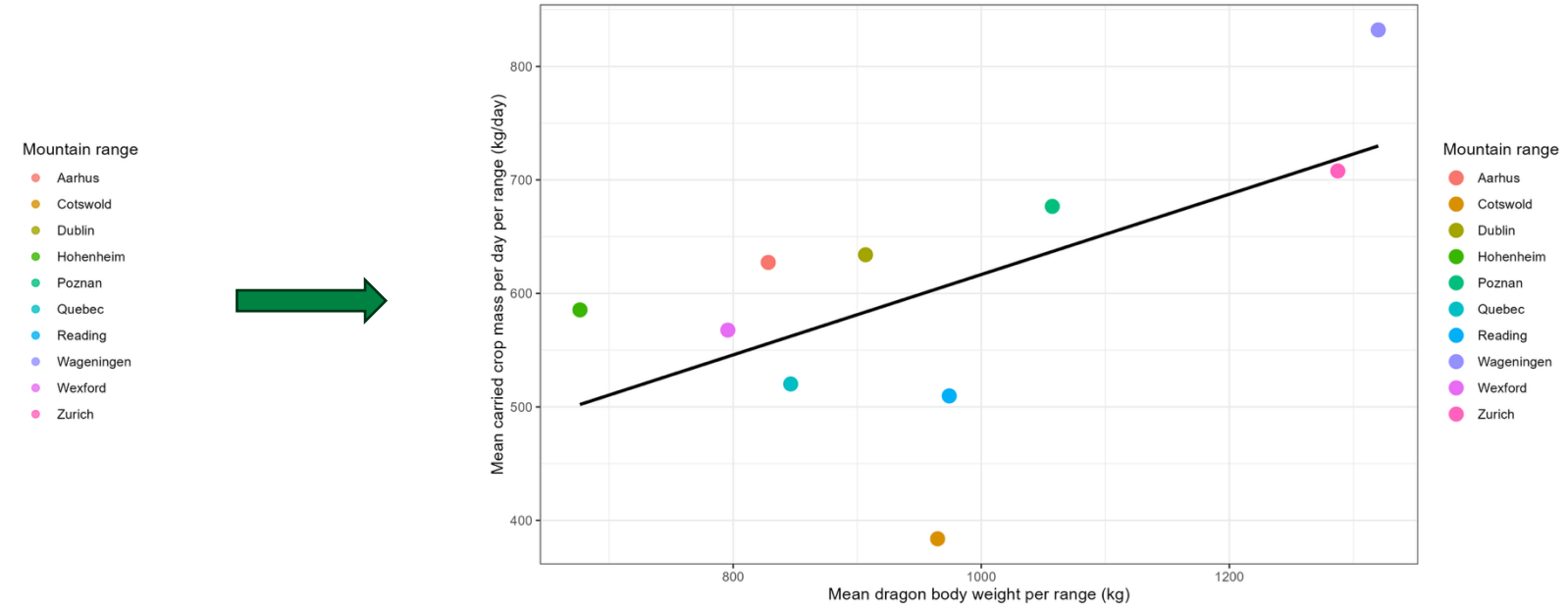
Calculate average

Dragon ID	Mountain range	Weight (kg)	Carried mass (kg)
1	Dublin	642	585
2	Dublin	567	755
3	Dublin	702	650
4	Dublin	690	670
…	…	…	…

Modelling strategies

Aggregate data





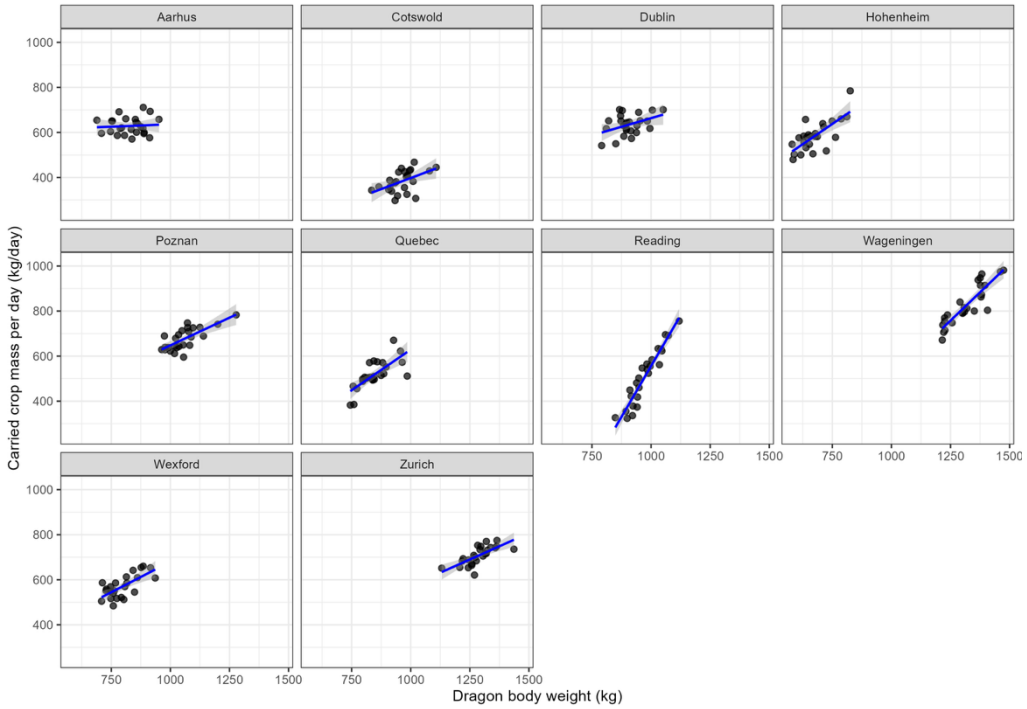
Calculate average

Dragon ID	Mountain range	Weight (kg)	Carried mass (kg)
1	Dublin	642	585
2	Dublin	567	755
3	Dublin	702	650
4	Dublin	690	670
...	...	...	...

Modelling strategies

Fit individual models

- Each group has its own intercept and slope
- Smaller sample size per model
- No single population-level effects
- No integrated group comparison



Modelling strategies

Mountain range as fixed block/factor

- Estimate only individual effects for each range
- Useful if we are only interested in these exact ranges
  - "If the mountain range cities hired me to compare performances between them"
  - Mountain range as factor variable ideal
  - With high number of groups, many coefficients to estimate
- Inference only on observed ranges, not a wider "population average"
  - "What happens if suddenly a new range appears?"
  - We couldn't estimate the performance of this range

$$y_i = \beta_0 + \beta_1 x_i + \alpha_j + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

$$i = 1, \dots, N = 250, \qquad j = 1, \dots, J = 10$$

Modelling strategies

What do we want?

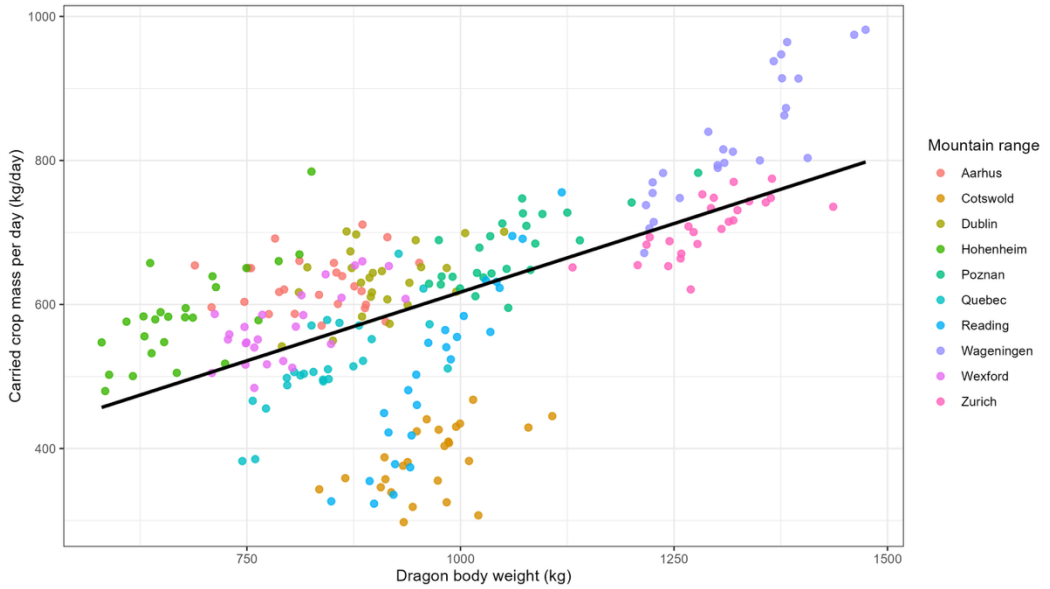
1. Keep all individual dragon-level observations
2. Estimate a population-level relationship between body weight and carried mass
3. Allow for non-independence within mountain ranges
4. Allow the body weight effect to differ between ranges
5. Avoid completely separated models
6. Quantify the mountain range-level variation in a parsimonious way

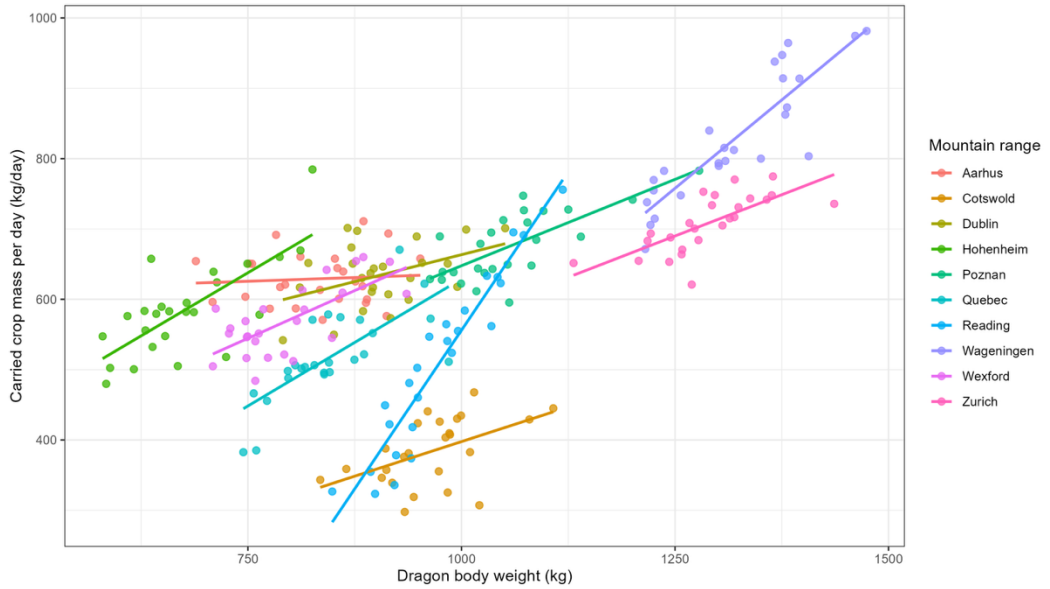
→ **This is the motivation for linear mixed-effects models**

Linear mixed-effects models

Concept

- Combination of population-level effects and individual group effects
- Mixed models combine fixed and random effects





Linear mixed-effects models

Concept

- Combination of population-level effects and individual group effects
- Mixed models combine fixed and random effects

Fixed effects	Random effects
Population-level effects	Group-level deviations
Average intercept/slope	Group-specific intercept/slope
Population-level relationship	Group-specific relationship

Observed outcome = Population-level fixed effect + Group-level deviation + Error

Linear mixed-effects models

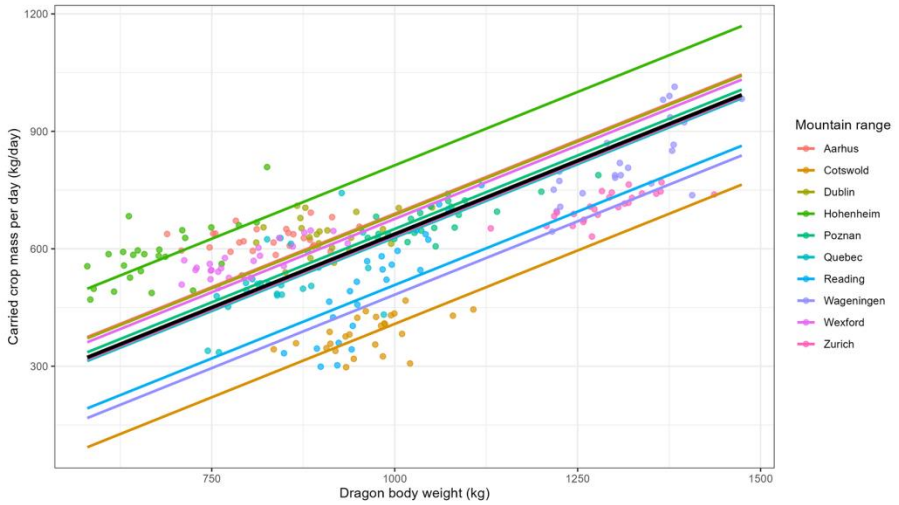
Random intercept

- Mixed models combine fixed and random effects
- Same average slope β_0, now with group specific intercept b_{0j}

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{ij} = (\beta_0 + b_{0j}) + \beta_1 x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$



Observed outcome = Population-level fixed effect + Group-level deviation + Error

Linear mixed-effects models

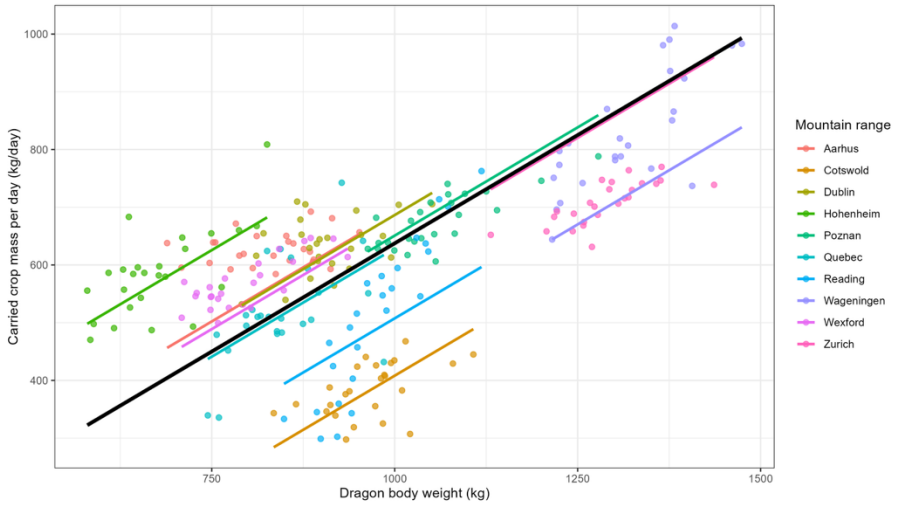
Random intercept

- Mixed models combine fixed and random effects
- Same average slope β_0, now with group specific intercept b_{0j}

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{ij} = (\beta_0 + b_{0j}) + \beta_1 x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$



Observed outcome = Population-level fixed effect + Group-level deviation + Error

Linear mixed-effects models

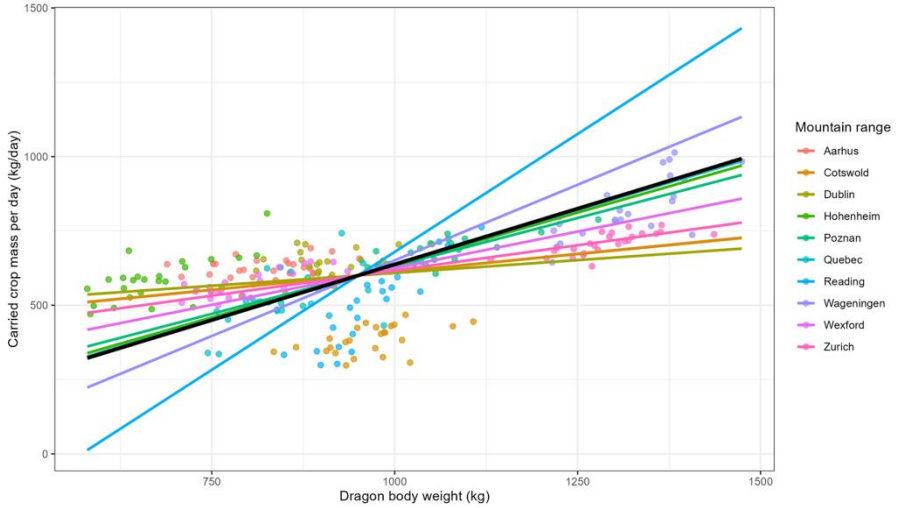
Random slope

- Mixed models combine fixed and random effects
- Same average slope β_1, now with group specific slope b_{1j}

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{ij} = \beta_0 + (\beta_1 + b_{1j}) x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$



Observed outcome = Population-level fixed effect + Group-level deviation + Error

Linear mixed-effects models

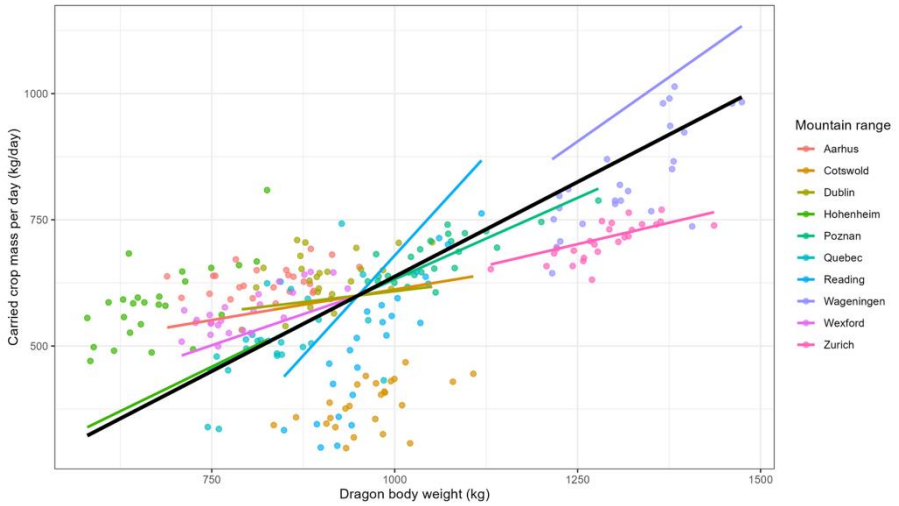
Random slope

- Mixed models combine fixed and random effects
- Same average slope β_1, now with group specific slope b_{1j}

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{ij} = \beta_0 + (\beta_1 + b_{1j}) x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$



Observed outcome = Population-level fixed effect + Group-level deviation + Error

Linear mixed-effects models

Random intercept + random slope

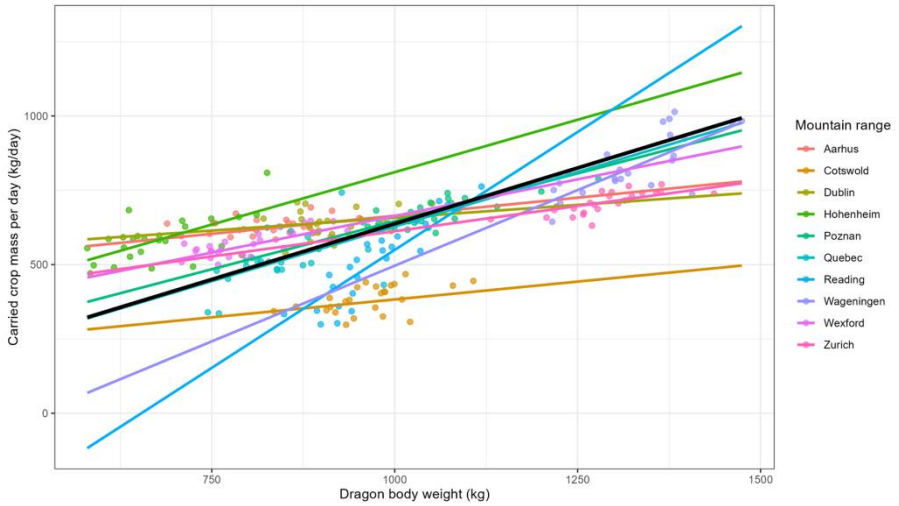
- Mixed models combine fixed and random effects
- A model with random intercept AND random slope

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j}) x_{ij} + \epsilon_{ij}$$

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{G})$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$



Observed outcome = Population-level fixed effect + Group-level deviation + Error

Linear mixed-effects models

Random intercept + random slope

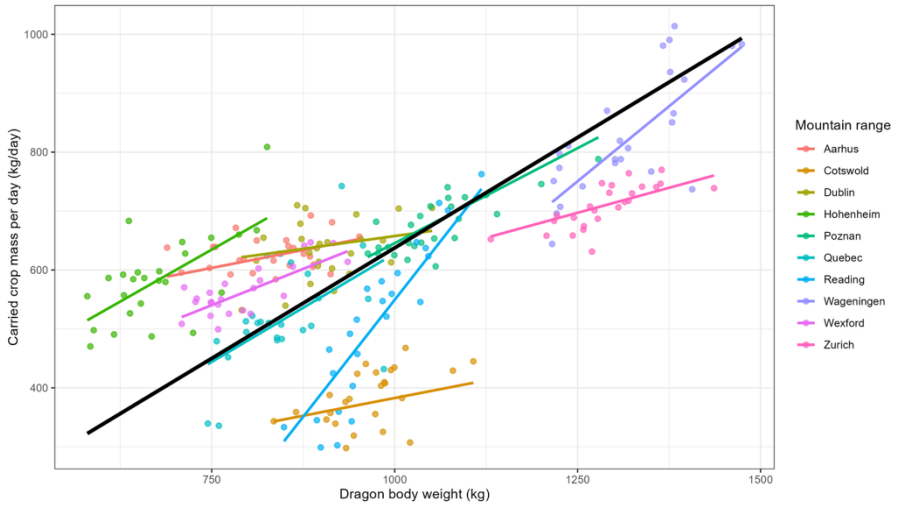
- Mixed models combine fixed and random effects
- A model with random intercept AND random slope

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j}) x_{ij} + \epsilon_{ij}$$

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{G})$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$



Observed outcome = Population-level fixed effect + Group-level deviation + Error

Linear mixed-effects models

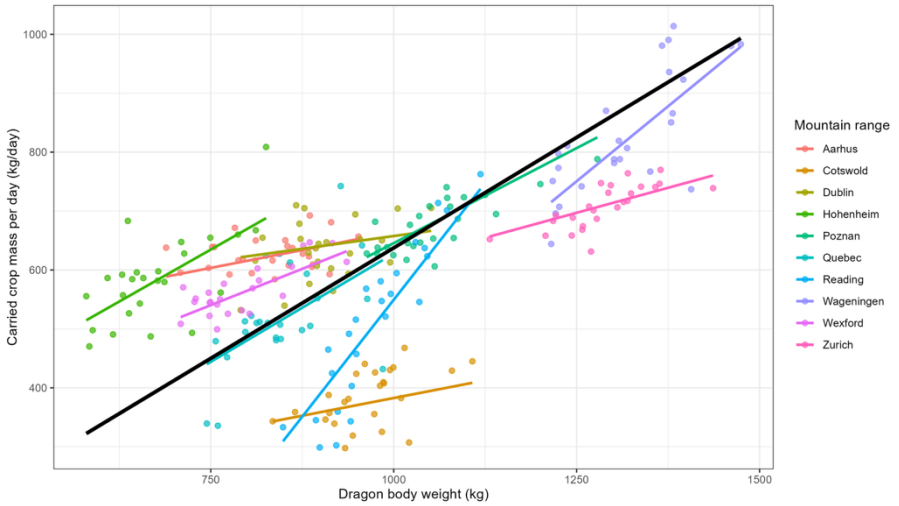
Random intercept + random slope

- Deviations around the population average (β_0, β_1)
- Model estimates how much groups vary around these population averages

$$y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{G}) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$



Observed outcome = Population-level fixed effect + Group-level deviation + Error

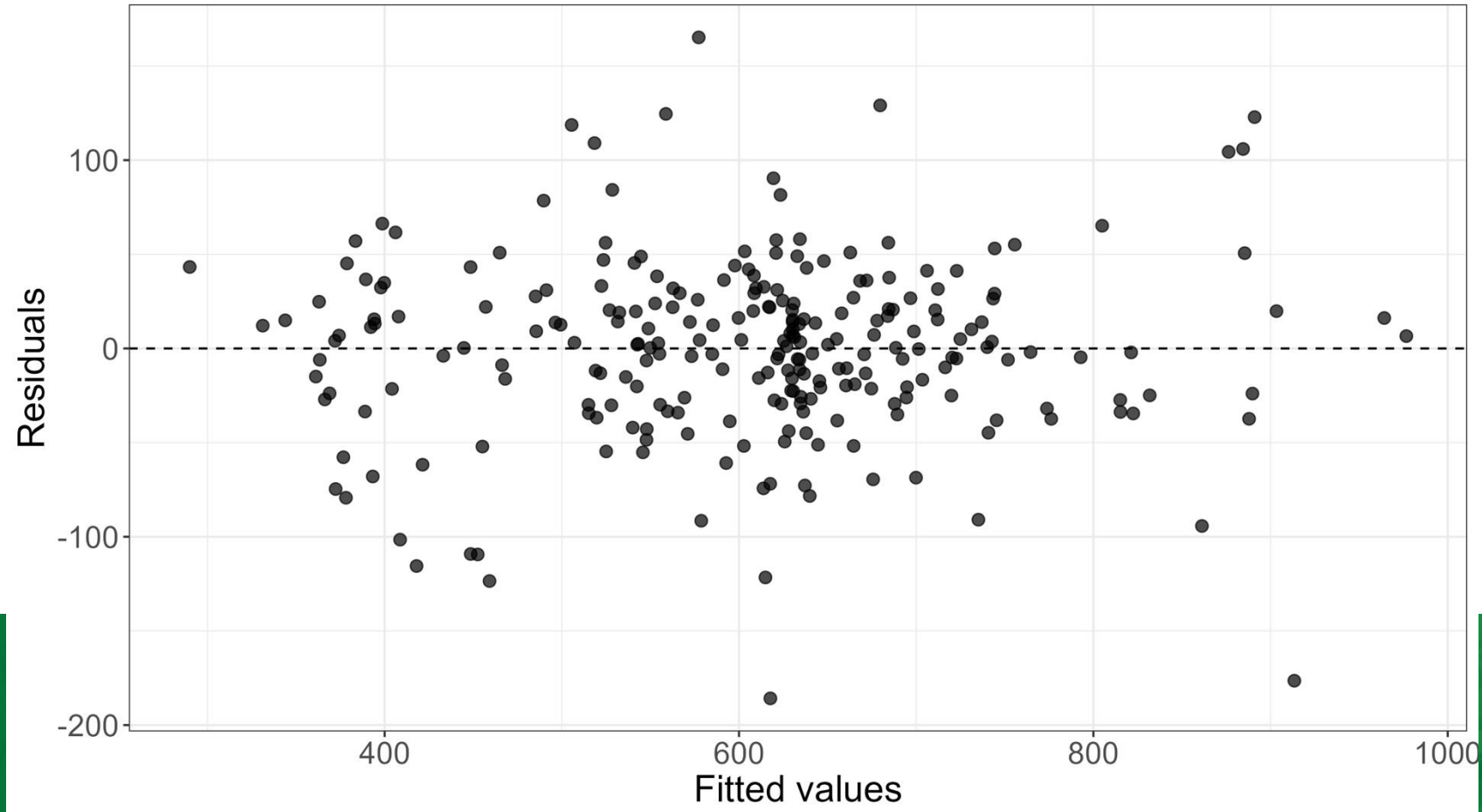
Linear mixed-effects models

Diagnostics

$$y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{G}) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$



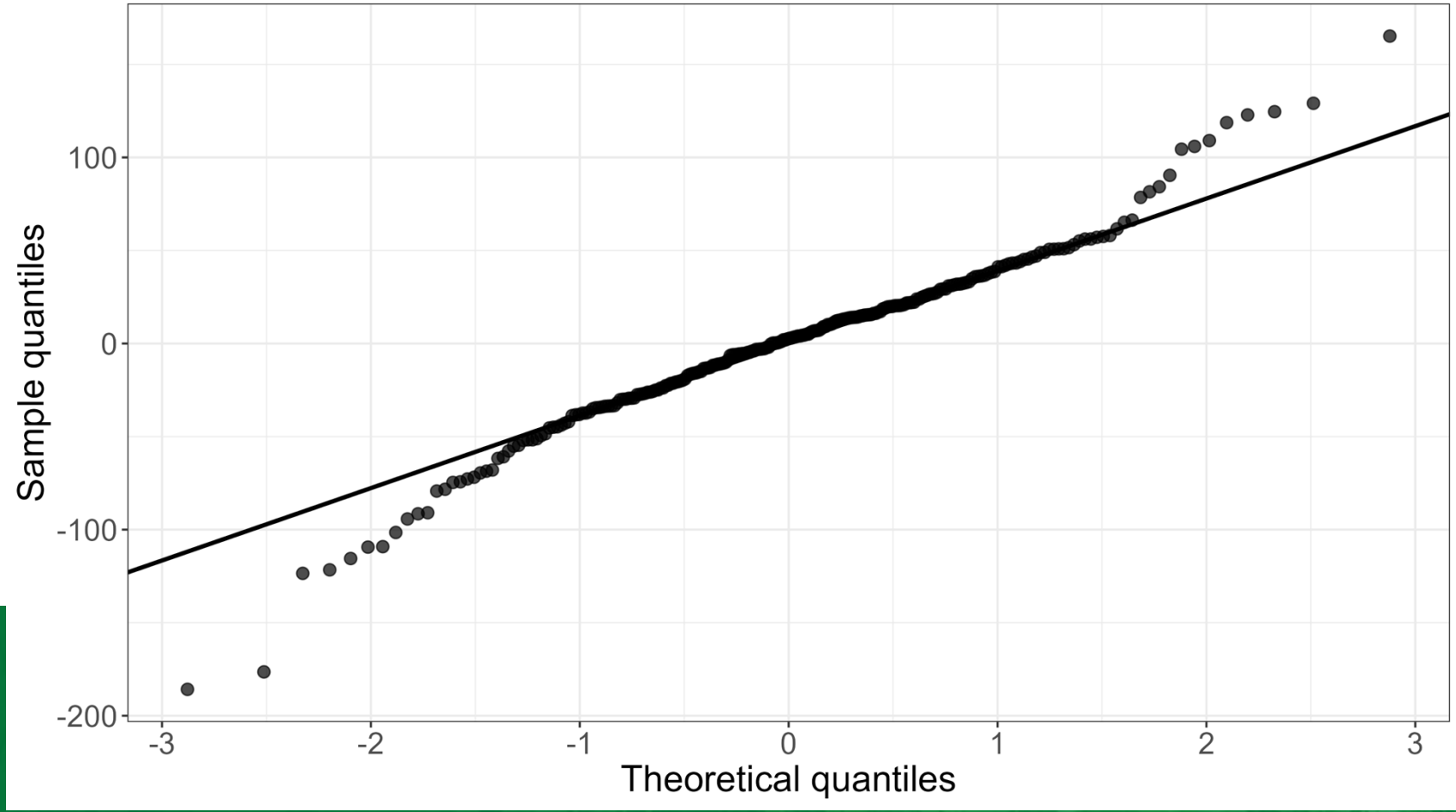
Linear mixed-effects models

Diagnostics

$$y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{G}) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$



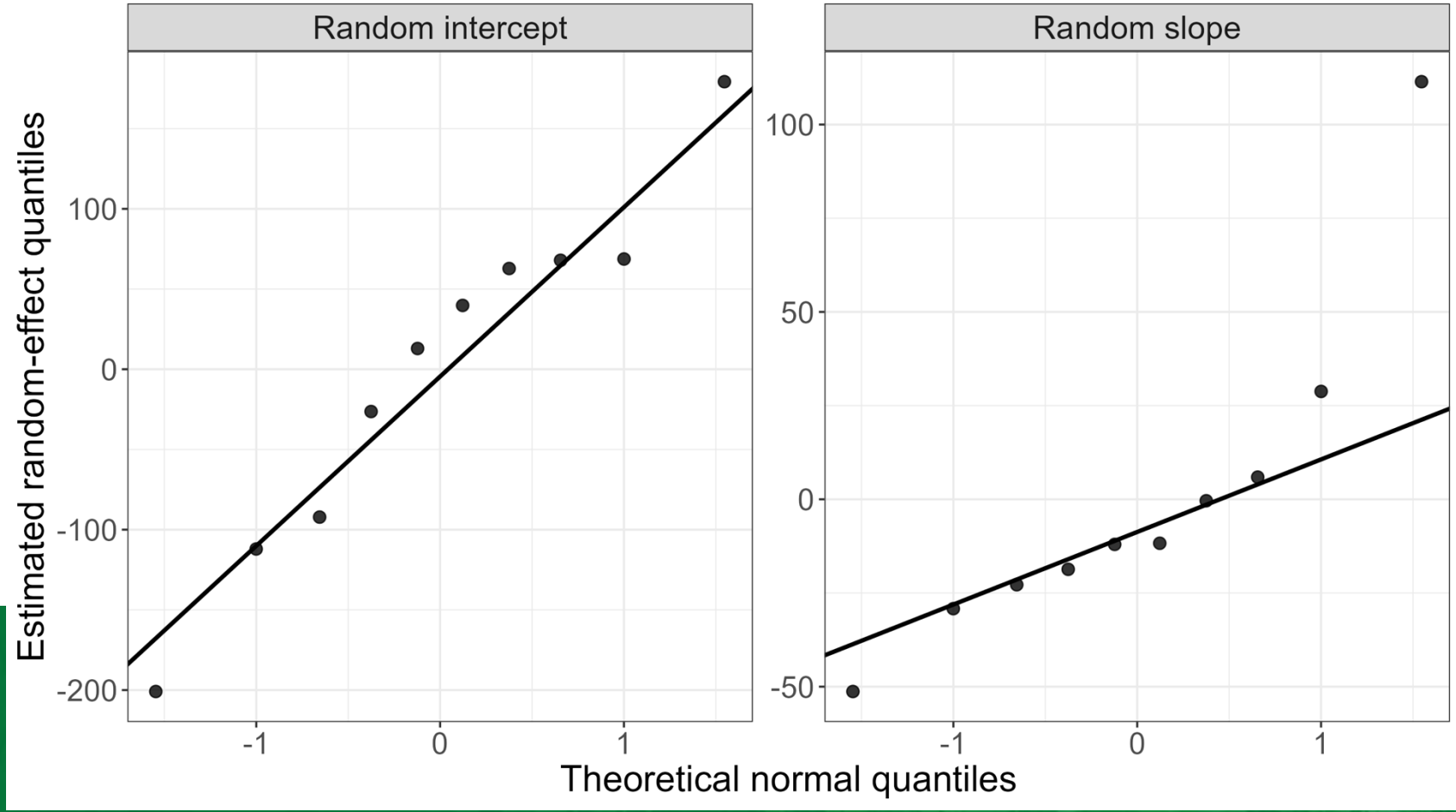
Linear mixed-effects models

Diagnostics

$$y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{G}) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$



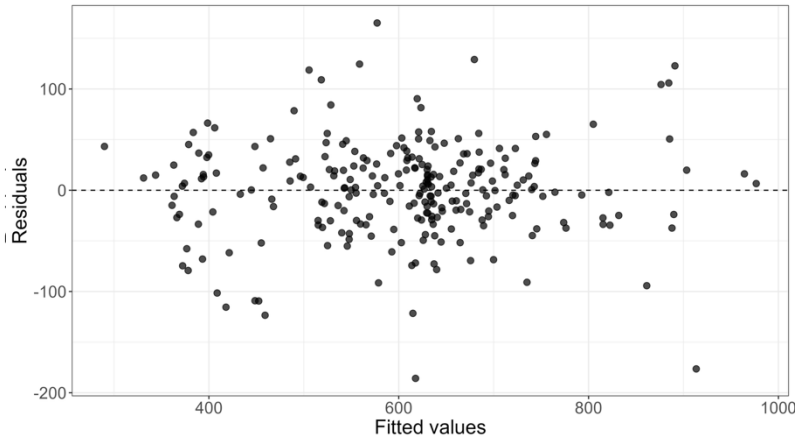
Linear mixed-effects models

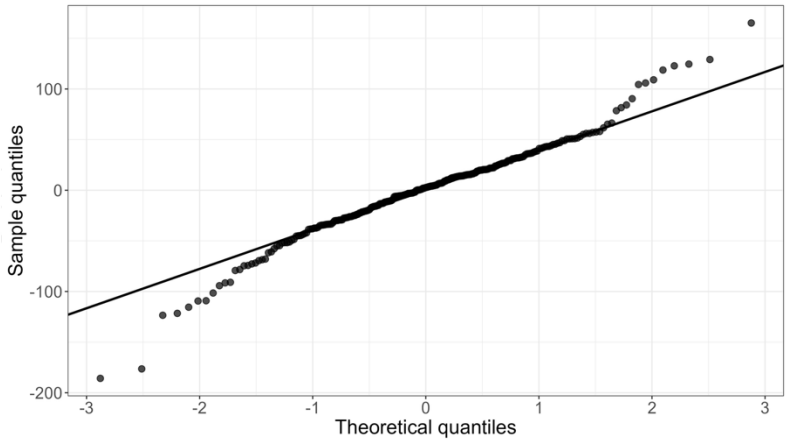
Diagnostics

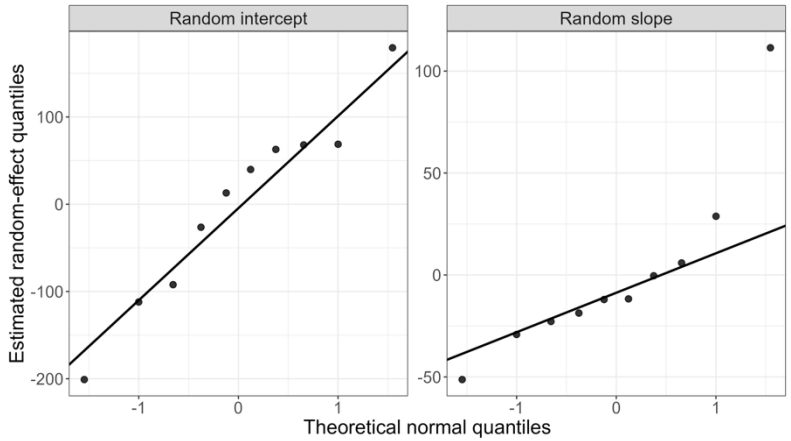
$$y_{ij} = (\beta_0 + b_{0j}) + (\beta_1 + b_{1j})x_{ij} + \epsilon_{ij}$$

$$i = 1, \ldots, N = 250, \qquad j = 1, \ldots, J = 10$$

$$\begin{pmatrix} b_{0j} \\ b_{1j} \end{pmatrix} \sim MVN(\mathbf{0}, \boldsymbol{G}) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$







Linear mixed-effects models

Mountain range as factor vs random effect

Mountain range as fixed block/factor	Mountain range as random block/effect
• The observed ranges are the main target of inference • We estimate named differences between ranges • Useful for comparing Dublin vs. Zurich vs. Aarhus etc. • We can't predict for a new range that hasn't been seen in the data • Often appropriate if groups were chosen deliberately because they are specifically of interest	• Observed ranges treated as samples from a broader population • We estimate the variation across ranges • Useful for generalising beyond the observed ranges • Allows to make predictions for new or unobserved ranges • Often appropriate if the groups/blocks/ranges represent possible levels from a broader population

→ The choice depends on the study design and the intended level of inference

Linear mixed-effects models

Summary

- Grouped data can violate assumption of independence
- Standard modelling approaches come with drawbacks
  - Different modelling strategies answer different questions
- Mixed-effects models keep all individual observations
  - Fixed effects represent population-average relationships
  - Allow predictors to vary by group
  - Random effects account for group-level deviations
  - Account for non-independence within groups


Thank you

Lab 1.1 – Exploring the data