Introduction to linear models

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

6 April 2020

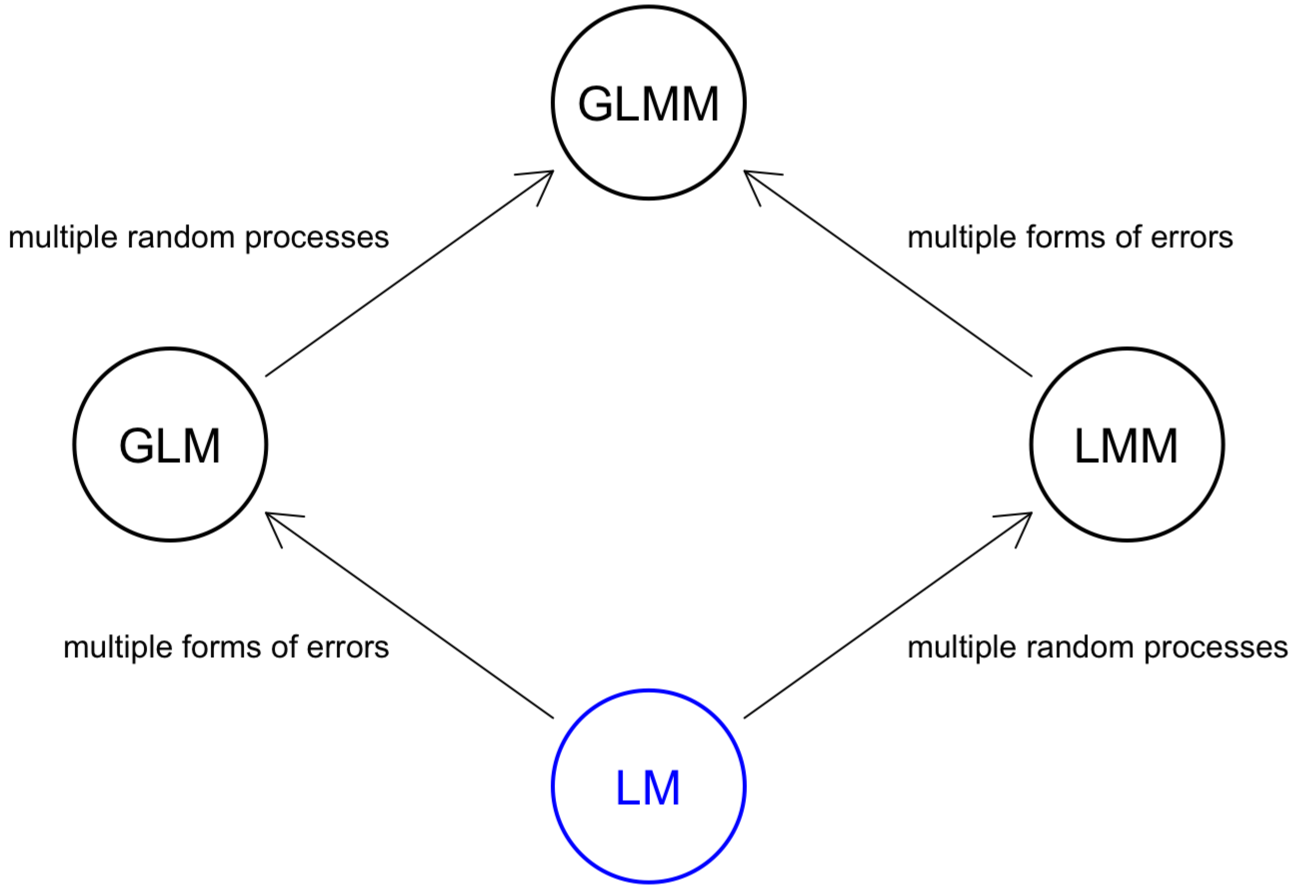
Goals for today

- Identify whether a model is linear in the predictors
- Recognize that linear models can approximate nonlinear functions
- Understand the difference between categorical and continuous models
- Recognize the difference between written and coded factors

Forms of linear models

Errors	Single random process	Multiple random processes
Normal	Linear Model (LM)	Linear Mixed Model (LMM)
Non-normal	Generalized Linear Model (GLM)	Generalized Linear Mixed Model (GLMM)

Forms of linear models



What is a linear model?

A relationship that defines a response variable as a linear function of one or more predictor variables

Which of these are linear models?

1) $y_i = \delta x_i$

2) $y_i = \alpha + \beta x_i$

3) $y_i = \alpha x_i^{\beta}$

4) $y_i = \alpha + \beta x_i + \gamma z_i$

5) $y_i = \alpha + \beta \frac{1}{x_i}$

6) $y_t = \mu + \phi(y_{t-1} - \mu)$

7) $y_i = (\alpha + x_i)\beta x_i$

8) $y_i = \frac{\alpha x_i}{1+\beta x_i}$

Which of these are linear models?

1) $y_i = \delta x_i$

2) $y_i = \alpha + \beta x_i$

3) $y_i = \alpha x_i^{\beta}$

4) $y_i = \alpha + \beta x_i + \gamma z_i$

5) $y_i = \alpha + \beta \frac{1}{x_i}$

6) $y_t = \mu + \phi(y_{t-1} - \mu)$

7) $y_i = (\alpha + x_i)\beta x_i$

8) $y_i = \frac{\alpha x_i}{1+\beta x_i}$

What is a linear model?

A relationship that defines a response variable as a linear function of one or more predictor variables

- characterized by a sum of terms, each of which is the product of a parameter and a single predictor

Is this a linear model?

$$y_i = \alpha(1 + \beta x_i)$$

Is this a linear model?

$$y_i = \alpha(1 + \beta x_i)$$

Yes, *if*

$$\begin{aligned} y_i &= \alpha(1 + \beta x_i) \\ &= \alpha + \alpha\beta x_i \\ &= \alpha + \gamma x_i \;\text{ with }\; \gamma = \alpha\beta \end{aligned}$$

What is a linear model?

A relationship that defines a response variable as a linear function of one or more predictor variables

- characterized by a sum of terms, each of which is the product of a parameter and a single predictor
- the predictor can be a transformed variable

Linear transformations

$$
\begin{aligned}
y_i &= \alpha + \beta x_i^2 \\
&\Downarrow \\
y_i &= \alpha + \beta z_i \\
z_i &= x_i^2
\end{aligned}
$$

Linear vs nonlinear models

There are only 2 forms of a linear model with 2 parameters

$$y_i = \alpha + \beta x_i$$

or

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i}$$

Linear vs nonlinear models

There are *many* forms of nonlinear models with 2 parameters

$$y_i = \alpha x_i^\beta$$

$$y_i = \alpha + x_i^\beta$$

$$y_i = \alpha^{\beta x_i}$$

$$y_i = \alpha + \beta \frac{1}{x}$$

$$\vdots$$

Linear vs nonlinear models

In linear models, effect sizes of different predictors are directly comparable

- intercept: units = response (eg, grams)
- slope: units = response per predictor (eg, grams per cm)

Linear vs nonlinear models

In linear models, effect sizes of different predictors are directly comparable

- intercept: units = response (eg, grams)
- slope: units = response per predictor (eg, grams per cm)

In nonlinear models, common inference tools (*p*-values, confidence intervals) may not be available

Locally linear models

If we reduce the scale (interval) enough, we can approximate a nonlinear function with a linear model

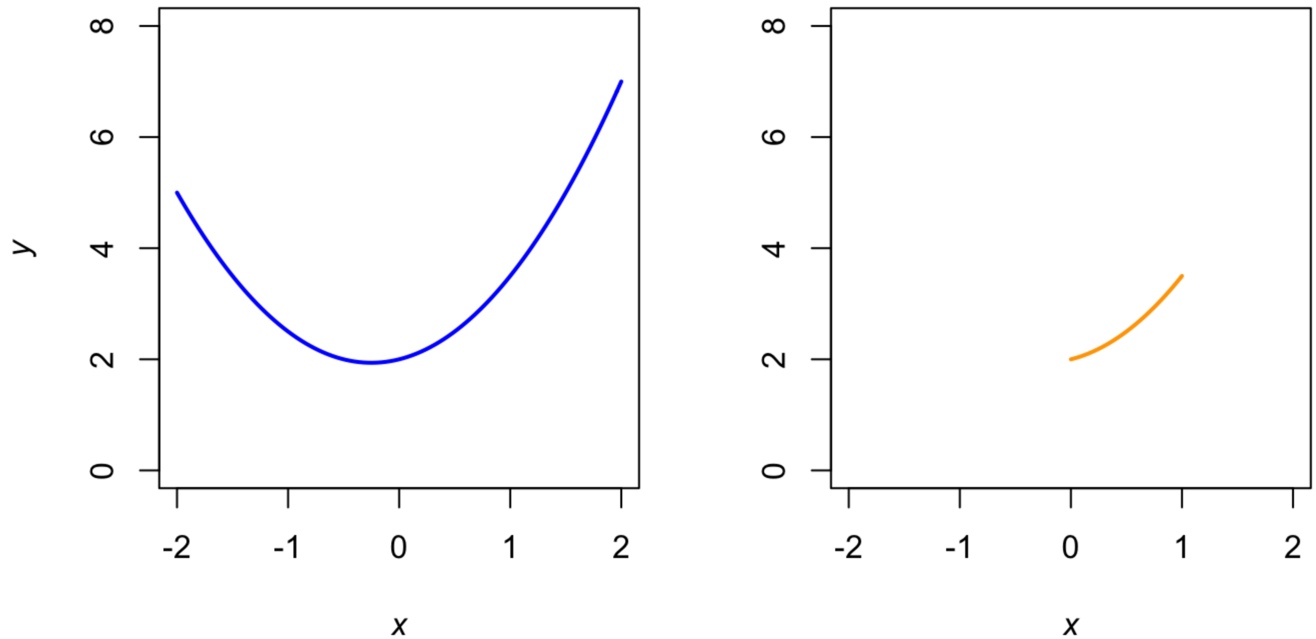
$$y = x^2$$

$$\Downarrow$$

$$\frac{dy}{dx} = 2x$$

Locally linear models

Consider the quadratic $y = \alpha + \beta x + x^2$



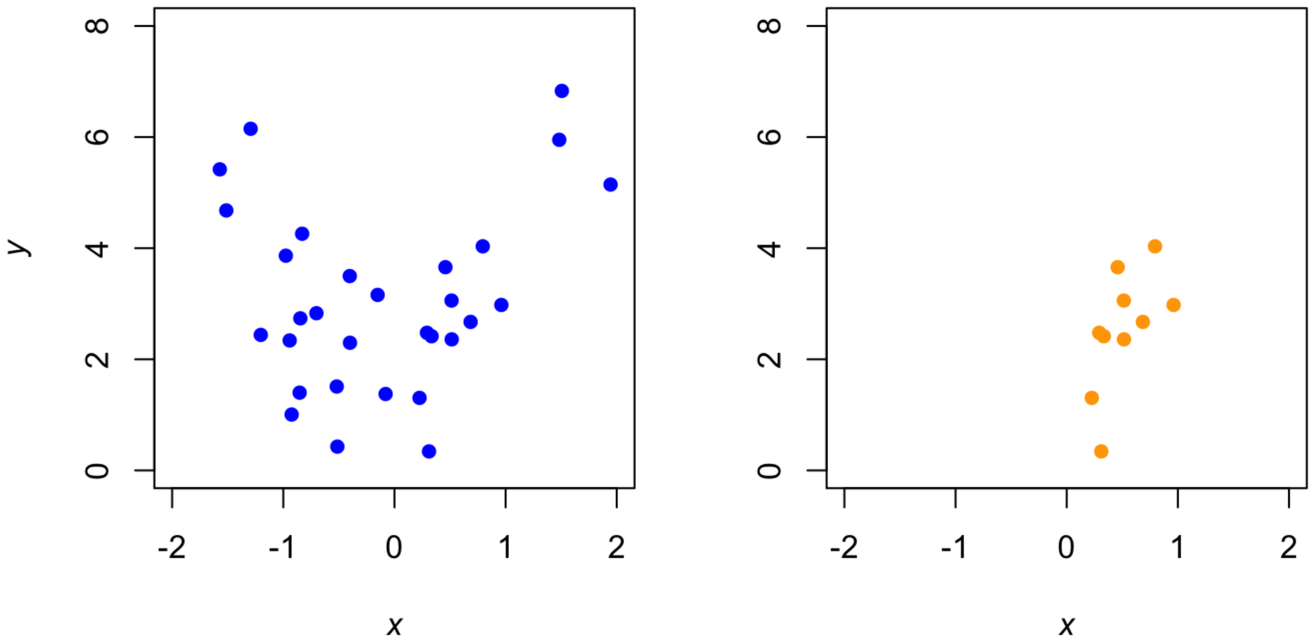
Locally linear models

A stochastic example with $y = \frac{1}{2} + 2x + x^2 + \epsilon_i$

```
set.seed(514)
nn <- 30
alpha <- 2
beta <- 1/2
eps <- rnorm(nn, 0, 1) ## errors ~ N(0,1)
x_all <- runif(nn, -2, 2)
y_all <- alpha + beta*x_all + x_all^2 + eps
x_loc <- x_all[x_all >= 0 & x_all <= 1]
y_loc <- y_all[x_all >= 0 & x_all <= 1]
```

Locally linear models

A stochastic example with $y = \frac{1}{2} + 2x + x^2 + \epsilon_i$



Linear model for size of fish

In **R**, we can use `lm()` to fit linear regression models

$$y_i = \alpha + \beta x_i + e_i$$

```
lm(y ~ x)
```

(notice that the intercept α is implicit here)

Linear model for size of fish

In **R**, we use `summary()` to get info about a fitted model

```
fitted_regr_model <- lm(L10_mass ~ L10_length)

summary(fitted_regr_model)
```

Locally linear models

```r
## model 1: full dataset
fit_1 <- lm(y_all ~ x_all)
summary(fit_1)
```

```
## 
## Call:
## lm(formula = y_all ~ x_all)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.8928 -0.9158 -0.2639  0.9593  3.4595 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   3.1293     0.3075  10.176 6.54e-11 ***
## x_all         0.3395     0.3339   1.017    0.318    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.669 on 28 degrees of freedom
## Multiple R-squared:  0.03559,    Adjusted R-squared:  0.001152 
## F-statistic: 1.033 on 1 and 28 DF,  p-value: 0.3181
```

Locally linear models

```
## model 2: "local" data
fit_2 <- lm(y_loc ~ x_loc)
summary(fit_2)
```

```
##
## Call:
## lm(formula = y_loc ~ x_loc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.6465 -0.4216  0.0882  0.5340  1.2668
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.1381     0.6993   1.627   0.1423
## x_loc         2.7334     1.2539   2.180   0.0609 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9015 on 8 degrees of freedom
## Multiple R-squared:  0.3727, Adjusted R-squared:  0.2942
## F-statistic: 4.752 on 1 and 8 DF,  p-value: 0.06087
```

Linear model for size of fish

In **R**, we use `coef()` to extract the intercept(s) and slope(s)

```
fitted_regr_model <- lm(y ~ x)

coef(fitted_regr_model)
```

Locally linear models

```
## intercept and slope for model 2
coef(fit_2)
```

```
## (Intercept)       x_loc
##    1.138064    2.733440
```

Locally linear models

```
## intercept and slope for model 2
coef(fit_2)
```

```
## (Intercept)       x_loc
##    1.138064    2.733440
```

True model: $y = \frac{1}{2} + 2x + x^2$

Estimate: $\hat{y} \approx 1.1 + 2.7x + 0x^2$

Linear models can be *good approximations* to nonlinear functions

QUESTIONS?

Common forms for linear models

A simple starting point

Data = (Deterministic part) + (Stochastic part)

Types of linear models

We classify linear models by the form of their deterministic part

Discrete predictor → ANalysis Of VAriance (ANOVA)

Continuous predictor → Regression

Both → ANalysis of COVAriance (ANCOVA)

Possible models for growth of fish

Model	Description
$\text{growth}_i = \alpha + \beta\text{species}_i + \epsilon_i$	1-way ANOVA
$\text{growth}_i = \alpha + \beta_{1,\text{species}} + \beta_{2,\text{tank}} + \epsilon_i$	2-way ANOVA
$\text{growth}_i = \alpha + \beta\text{ration}_i + \epsilon_i$	simple linear regression
$\text{growth}_i = \alpha + \beta_1\text{ration}_i + \beta_2\text{temperature}_i + \epsilon_i$	multiple regression
$\text{growth}_i = \alpha + \beta_{1,\text{species}} + \beta_2\text{ration}_i + \epsilon_i$	ANCOVA

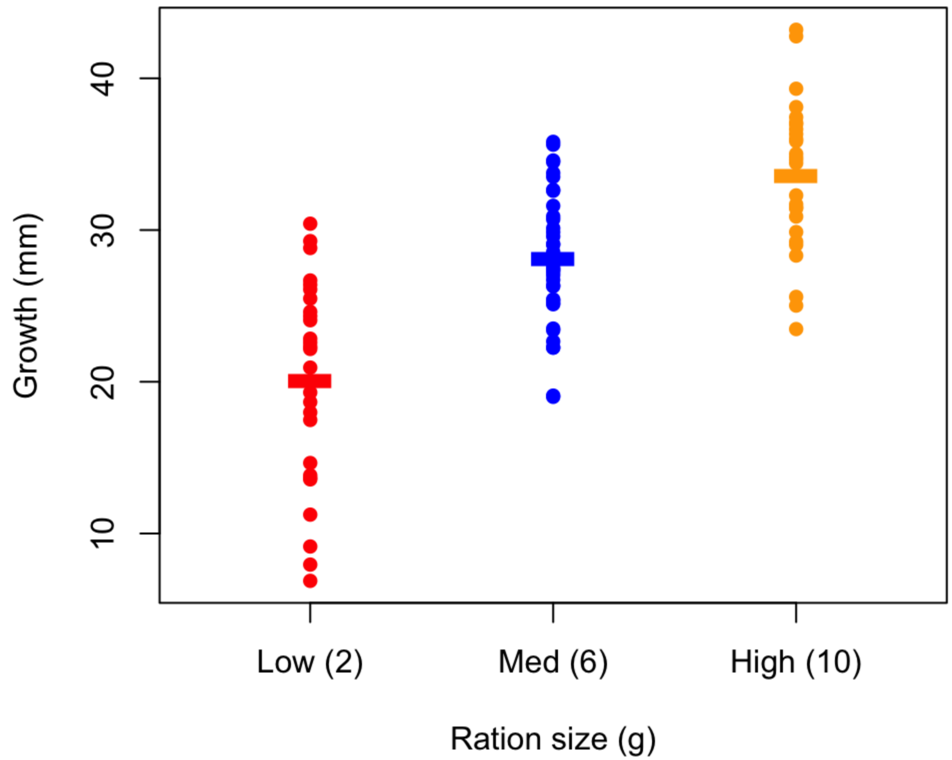
Example

Fish growth during an experiment

- A biologist at the WA Dept of Fish & Wildlife contacts you for help with an experiment
- She wants to know how growth of hatchery salmon is affected by their ration size
- She sends you a spreadsheet with 2 cols:

  1. fish growth (mm)
  2. ration size (2g, 4g, 6g)

ANOVA model

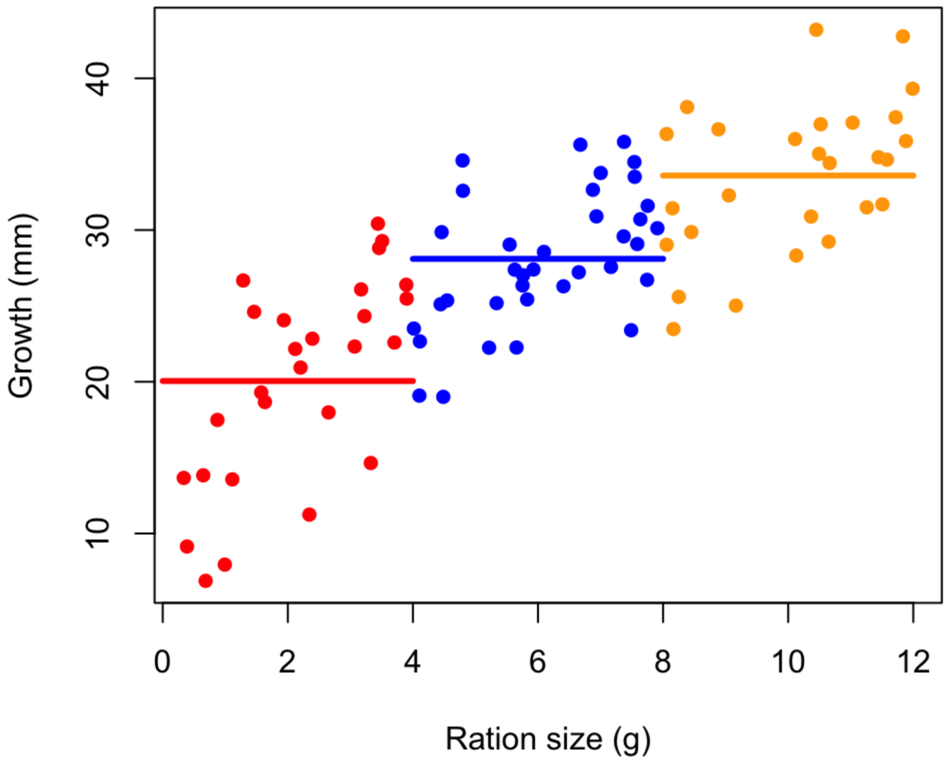


$$\text{growth}_i = \alpha + \beta_{\text{ration}} + \epsilon_i$$

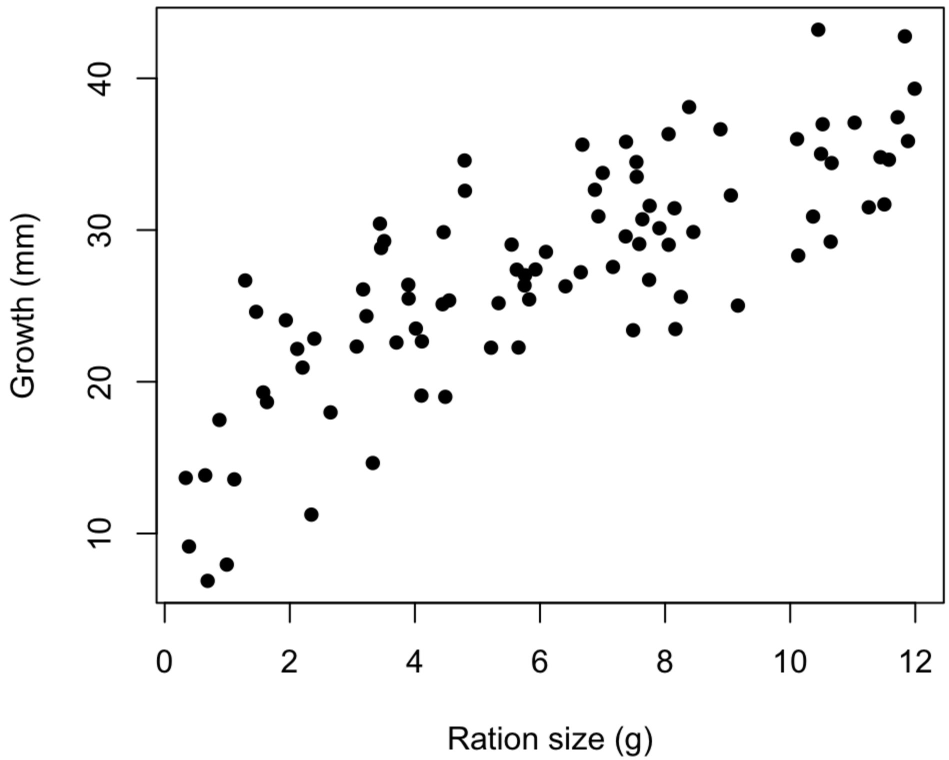
More info arrives

It turns out that targeting the exact ration is hard, but they know how much each fish ate during the trial

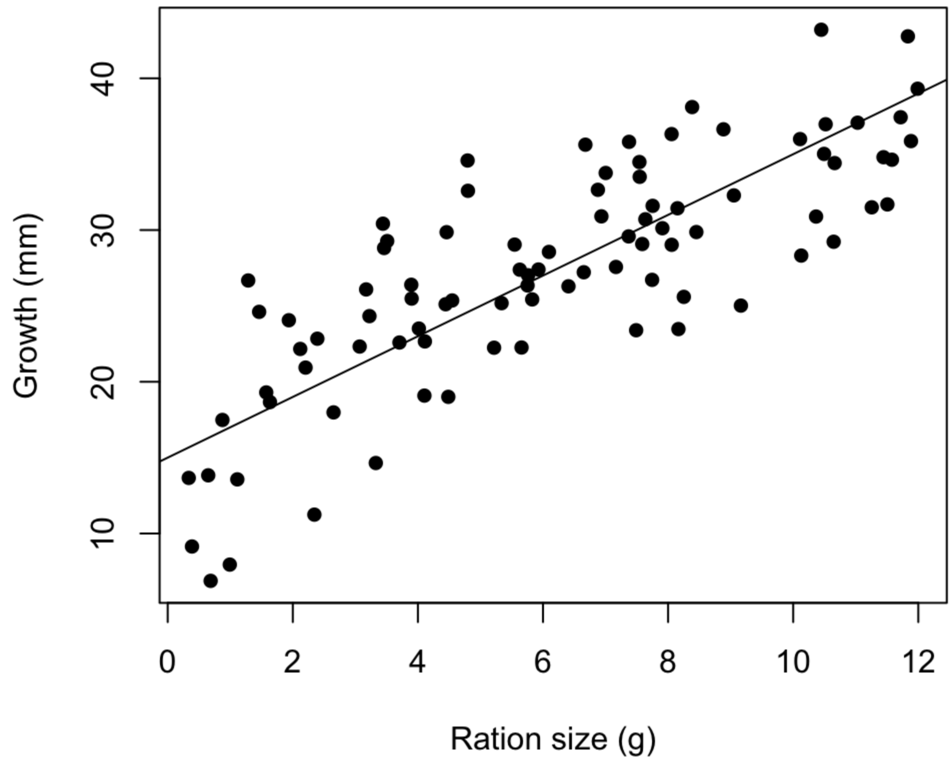
Continuous predictor



Continuous predictor



Linear regression

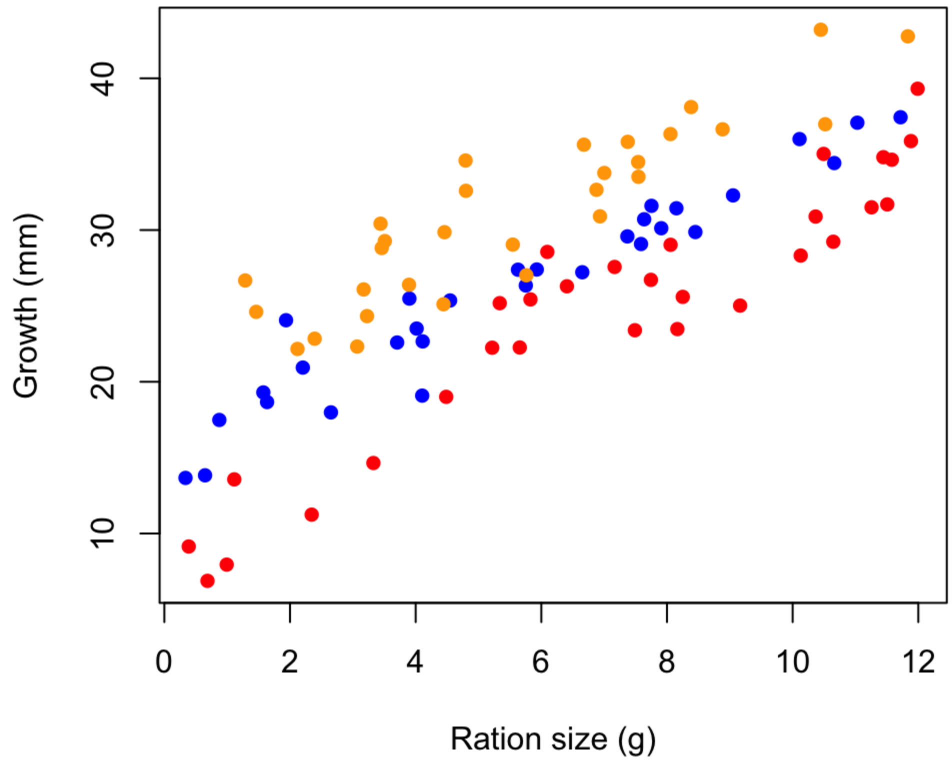


$$\text{growth}_i = \alpha + \beta \text{ration}_i + \epsilon_i$$

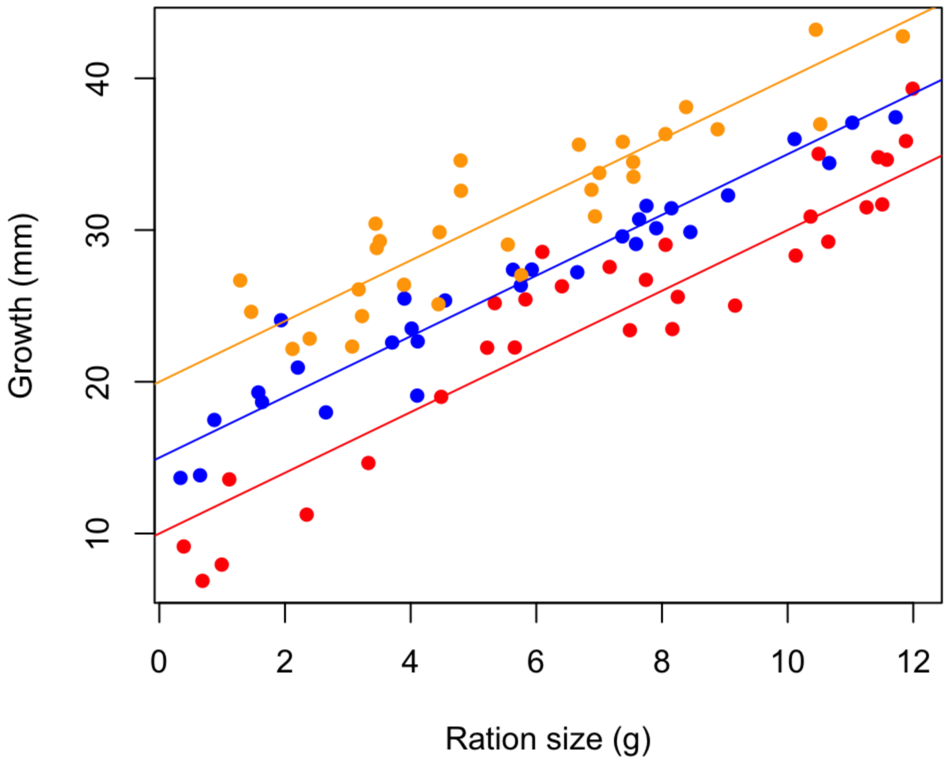
More info arrives

It also turns out that there are 3 lineages of fish in the trials

Continuous & discrete predictors



ANCOVA



$$\text{growth}_i = \alpha + \beta_{1,\text{lineage}} + \beta_2 \text{ration}_i + \epsilon_i$$

Notation for categorical effects

Here we have specified categorical effects in AN(C)OVA models as discrete parameters

For example, for a one-way ANOVA with 3 factors

$$y_i = \alpha + \beta_j + \epsilon_i$$

the definition of β_j is

$$\beta_j = \begin{cases} \beta_1 \text{ if factor 1} \\ \beta_2 \text{ if factor 2} \\ \beta_3 \text{ if factor 3} \end{cases}$$

Notation for categorical effects

In practice, we will use a combination of -1/0/1 predictors, so our model becomes

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \epsilon_i$$

and each of the $x_{j,i}$ indicates whether the i^{th} observation was assigned factor j

(We'll visit this again when we discuss design matrices)