Design matrices for models

Analysis of Ecological and Environmental Data

QERM 514

Mark Scheuerell

20 April 2020

Goals for today

- Understand how to create design matrices for use in linear models
- Recognize the different coding schemes for factor models
- See how to use `model.matrix()` for creating & extracting design matrices

Models in matrix form

Recall the matrix form for our linear models, where

$$
\begin{aligned}
\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\
\mathbf{e} &\sim \text{MVN}(\mathbf{0}, \boldsymbol{\Sigma})
\end{aligned}
$$

Models in matrix form

Let's write out this model in more detail

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{n,1} \\ 1 & x_{1,2} & \cdots & x_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{n,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

The columns in $\mathbf{X}$ define the *design* of the analysis

Ordinary least squares

Also recall that we can use $\mathbf{X}$ to solve for $\hat{\mathbf{y}}$

$$
\begin{aligned}
\hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} \\
&= \mathbf{X}\left((\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\right) \\
&= \underbrace{\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}_{\mathbf{H}}\mathbf{y} \\
&= \mathbf{H}\mathbf{y}
\end{aligned}
$$

Understanding the form of $\mathbf{X}$ is critical to our inference

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
$\text{growth}_i = \beta_0 + \beta_{1,\text{species}} + \epsilon_i$	1-way ANOVA
$\text{growth}_i = \beta_0 + \beta_{1,\text{species}} + \beta_{2,\text{tank}} + \epsilon_i$	2-way ANOVA
$\text{growth}_i = \beta_0 + \beta_1 \text{ration}_i + \epsilon_i$	simple linear regression
$\text{growth}_i = \beta_0 + \beta_1 \text{ration}_i + \beta_2 \text{temperature}_i + \epsilon_i$	multiple regression
$\text{growth}_i = \beta_0 + \beta_{1,\text{species}} + \beta_2 \text{ration}_i + \epsilon_i$	ANCOVA

Defining models with $\mathbf{X}$

Mean only

What would $\mathbf{X}$ look like for a simple model of the data $\mathbf{y}$ that included a mean only?

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$$

Defining models with $\mathbf{X}$

Mean only

Let's start by rewriting our model as

$$
\begin{aligned}
\mathbf{y} &= \boldsymbol{\beta}_0 + \mathbf{e} \\
&= \begin{bmatrix} \boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_0 \\ \vdots \\ \boldsymbol{\beta}_0 \end{bmatrix} + \mathbf{e}
\end{aligned}
$$

Defining models with $\mathbf{X}$

Mean only

$$
\begin{aligned}
\mathbf{y} &= \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \boldsymbol{\beta}_0 + \mathbf{e} \\
&= \mathbf{X} \boldsymbol{\beta} + \mathbf{e}
\end{aligned}
$$

with $\mathbf{X} = [1 \ 1 \cdots 1]^\top$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_0]$

Defining models with $\mathbf{X}$

Regression

What would $\mathbf{X}$ look like for a regression model with 2 predictors?

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + e_i$$
$$\Downarrow ?$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Defining models with $\mathbf{X}$

Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} \\ 1 & x_{1,2} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Defining models with $\mathbf{X}$

Regression

What would $\mathbf{X}$ look like for model with an intercept and linear increase over time t?

$$
\begin{gathered}
y_t = \beta_0 + \beta_1 t + e_t \\
\Downarrow ? \\
\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}
\end{gathered}
$$

Defining models with $\mathbf{X}$

Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & ? \\ 1 & ? \\ \vdots & \vdots \\ 1 & ? \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Defining models with $\mathbf{X}$

Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

$$\Downarrow$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \vdots \\ e_n \end{bmatrix}$$

Defining models with $\mathbf{X}$

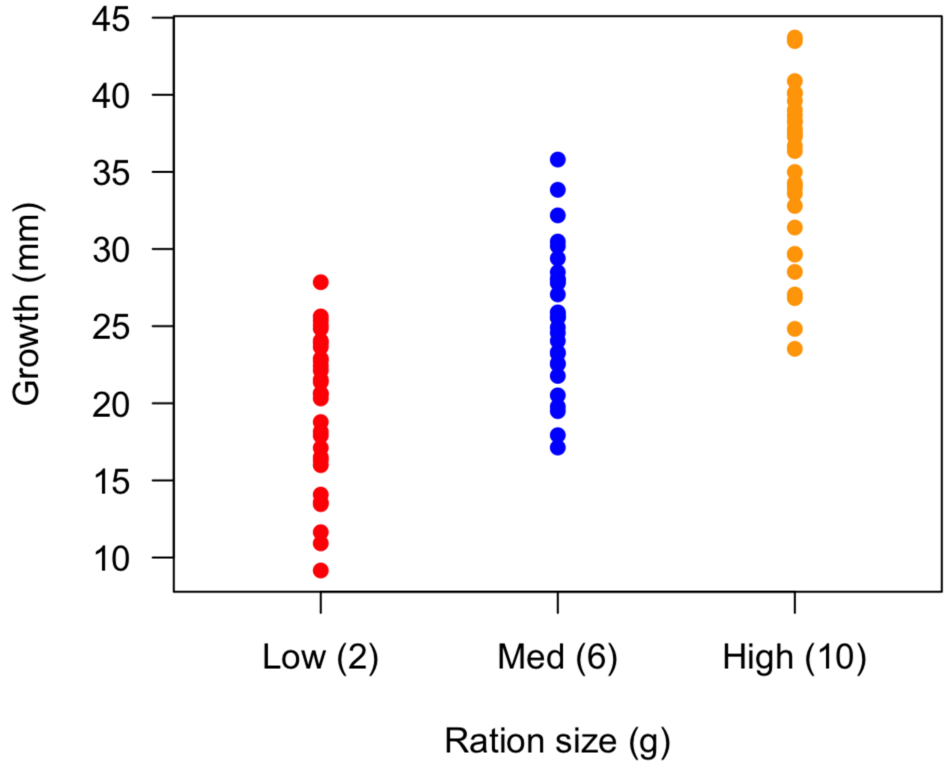
Analysis of variance (ANOVA)

ANOVA was popularized by Ronald Fisher ~100 years ago when he was studying the variance of genetic traits among commercial crops

ANOVA is used to analyze *differences among group means*

Comparing group means

Recall our analysis of fish growth as a function of ration



Defining models with $\mathbf{X}$

ANOVA

Here we want to know if the mean growth of fish varies among the 3 ration sizes

$$\bar{\boldsymbol{g}}_{\text{ration}_1} \stackrel{?}{=} \bar{\boldsymbol{g}}_{\text{ration}_2} \stackrel{?}{=} \bar{\boldsymbol{g}}_{\text{ration}_3}$$

How would we write the model for this?

Defining models with $\mathbf{X}$

ANOVA

Our model for an observation y_i is something like

$$y_i = \mu_i + e_i$$

$$\mu_i = \begin{cases} \mu_1 & \text{if fed ration 1} \\ \mu_2 & \text{if fed ration 2} \\ \mu_3 & \text{if fed ration 3} \end{cases}$$

Defining models with $\mathbf{X}$

ANOVA

We can use binary 0/1 coding to represent if/then constructs

$$y_i = \mu_1 x_{1,i} + \mu_2 x_{2,i} + \mu_3 x_{3,i} + e_i$$

$$\begin{aligned} x_{1,i} &= 1 \text{ if fed ration 1 and 0 otherwise} \\ x_{2,i} &= 1 \text{ if fed ration 2 and 0 otherwise} \\ x_{3,i} &= 1 \text{ if fed ration 3 and 0 otherwise} \end{aligned}$$

Defining models with $\mathbf{X}$

ANOVA

How would we specify the model matrix $\mathbf{X}$ for this?

Defining models with $\mathbf{X}$

ANOVA

Let's rewrite our model as

$$y_i = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i$$
$$\Downarrow$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

Defining models with $\mathbf{X}$

ANOVA

And define $\mathbf{X}$ as

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_{1,1} & \boldsymbol{x}_{2,1} & \boldsymbol{x}_{3,1} \\ \boldsymbol{x}_{1,2} & \boldsymbol{x}_{2,2} & \boldsymbol{x}_{3,2} \\ \vdots & \vdots & \vdots \\ \boldsymbol{x}_{1,n} & \boldsymbol{x}_{2,n} & \boldsymbol{x}_{3,n} \end{bmatrix}$$

Defining models with $\mathbf{X}$

Let's now re-order all of the observations into their groups

$$\mathbf{y} = \begin{bmatrix} y_{1,1} \\ \vdots \\ y_{1,j_1} \\ \hline y_{2,1} \\ \vdots \\ y_{2,j_2} \\ \hline y_{3,1} \\ \vdots \\ y_{3,j_3} \end{bmatrix} \text{ with } j_1 + j_2 + j_3 = n$$

Defining models with $\mathbf{X}$

We can then define $\mathbf{X}$ and $\boldsymbol{\beta}$ as

$$\mathbf{X} = \left[\begin{array}{ccc} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{array}\right] \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Defining models with $\mathbf{X}$

ANOVA

Here are the mean growth rates of our 3 groups of fish

$\bar{y}_{j=1} = \beta_1 = 19.6$

$\bar{y}_{j=2} = \beta_2 = 25.6$

$\bar{y}_{j=3} = \beta_3 = 35$

Defining models with $\mathbf{X}$

ANOVA

And here are the results of our ANOVA model

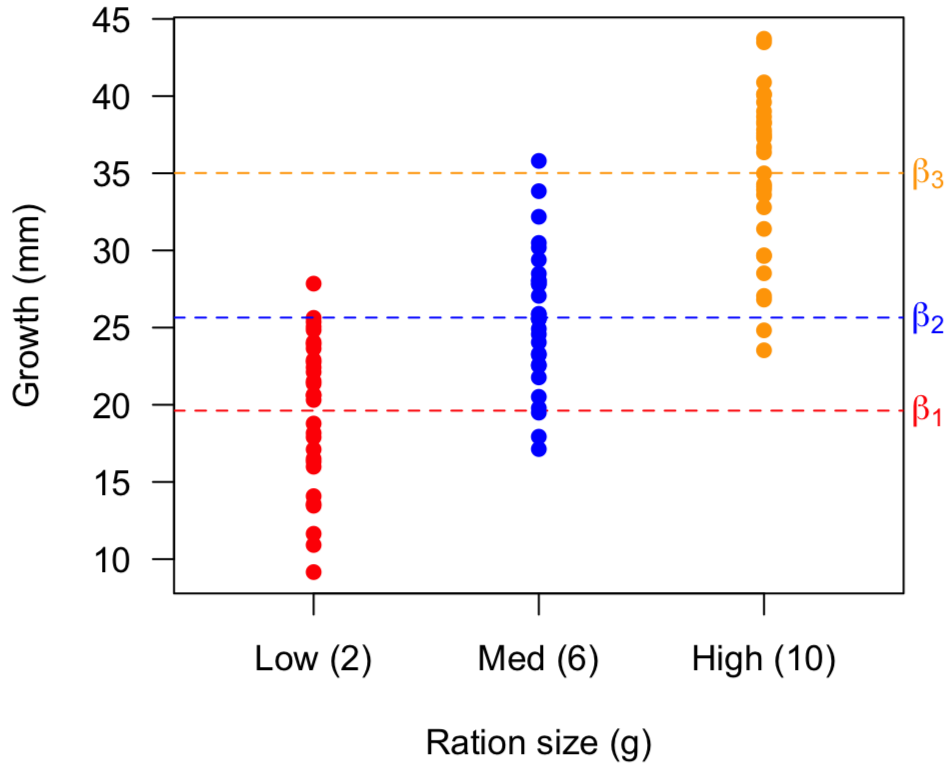
```
## fit ANOVA w/ `- 1` to remove intercept
m1 <- lm(yy ~ ration - 1)
coef(m1)
```

```
## ration_1 ration_2 ration_3
## 19.62001 25.64846 35.01523
```

This confirms that we have fit a model of means

Defining models with $\mathbf{X}$

ANOVA



Defining models with $\mathbf{X}$

ANOVA

Suppose we wanted to reframe our model to instead include the effect of ration relative to the overall mean growth rate (μ)

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

and calculate the groups means as

$$\begin{aligned}
\bar{y}_{j=1} &= \mu + \beta_1 \\
\bar{y}_{j=2} &= \mu + \beta_2 \\
\bar{y}_{j=3} &= \mu + \beta_3
\end{aligned}$$

Defining models with $\mathbf{X}$

We would then define $\mathbf{X}$ and $\boldsymbol{\beta}$ as

$$\mathbf{X} = \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{array}\right] \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 \end{bmatrix}$$

Defining models with $\mathbf{X}$

ANOVA

And here are the results of our ANOVA model

```
## design matrix
X <- cbind(rep(1,nn*pp), ration)
## fit ANOVA w/ `- 1` to remove intercept
m2 <- lm(yy ~ X - 1)
coef(m2)
```

```
##          X        X_1        X_2        X_3
##  35.015235 -15.395221  -9.366774         NA
```

Wait–what happened here?!

Defining models with $\mathbf{X}$

Can you spot the problem in our design matrix?

$$
\mathbf{X} = \begin{bmatrix}
1 & 1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 1 & 0 & 0 \\
\hline
1 & 0 & 1 & 0 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 0 & 1 & 0 \\
\hline
1 & 0 & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots \\
1 & 0 & 0 & 1
\end{bmatrix}
$$

Defining models with $\mathbf{X}$

ANOVA

```
## solve for beta by hand
beta <- solve(t(X) %*% X) %*% t(X) %*% yy
```

```
## Error in solve.default(t(X) %*% X) :
##   system is computationally singular: reciprocal condition number
```

Defining models with $\mathbf{X}$

$\mathbf{X}$ is not *full rank* ($\mathbf{X}_{(\cdot 1)} = \mathbf{X}_{(\cdot 2)} + \mathbf{X}_{(\cdot 3)} + \mathbf{X}_{(\cdot 4)}$)

$$\mathbf{X} = \left[\begin{array}{cccc} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{array}\right]$$

Defining models with $\mathbf{X}$

ANOVA

Let's think about our model again

$$\boldsymbol{y_i} = \boldsymbol{\mu} + \boldsymbol{\beta}_1 \boldsymbol{x}_{1,i} + \boldsymbol{\beta}_2 \boldsymbol{x}_{2,i} + \boldsymbol{\beta}_3 \boldsymbol{x}_{3,i} + \boldsymbol{e_i}$$

where we want the group means to be

$$\begin{aligned} \bar{\boldsymbol{y}}_{j=1} &= \boldsymbol{\mu} + \boldsymbol{\beta}_1 \\ \bar{\boldsymbol{y}}_{j=2} &= \boldsymbol{\mu} + \boldsymbol{\beta}_2 \\ \bar{\boldsymbol{y}}_{j=3} &= \boldsymbol{\mu} + \boldsymbol{\beta}_3 \end{aligned}$$

Defining models with $\mathbf{X}$

ANOVA

Consider the overall mean of $\mathbf{y}$ in terms of the group means

$$\bar{y} = \frac{\bar{y}_{j=1} + \bar{y}_{j=2} + \bar{y}_{j=3}}{3}$$

Defining models with $\mathbf{X}$

ANOVA

Consider the overall mean of $\mathbf{y}$ in terms of the group means

$$\bar{y} = \frac{\bar{y}_{j=1} + \bar{y}_{j=2} + \bar{y}_{j=3}}{3}$$

$$\Downarrow$$

$$\mu = \frac{(\mu + \beta_1) + (\mu + \beta_2) + (\mu + \beta_3)}{3}$$

$$\Downarrow$$

$$\beta_1 + \beta_2 + \beta_3 = 0$$

Defining models with $\mathbf{X}$

ANOVA

Now we can rewrite our model as

$$y_i = \mu + \beta_1 x_{1,i} + \beta_2 x_{2,i} + (\text{-}\beta_1 + \text{-}\beta_2) x_{3,i} + e_i$$

and calculate the group means as

$$\begin{aligned}
\bar{y}_{j=1} &= \mu + \beta_1 \\
\bar{y}_{j=2} &= \mu + \beta_2 \\
\bar{y}_{j=3} &= \mu - (\beta_1 + \beta_2)
\end{aligned}$$

Defining models with $\mathbf{X}$

We would then define $\mathbf{X}$ and $\boldsymbol{\beta}$ as

$$\mathbf{X} = \left[\begin{array}{ccc} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ \hline 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & -1 & -1 \end{array}\right] \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$$

Defining models with $\mathbf{X}$

ANOVA

```
## empty design matrix
XX <- matrix(NA, nn*pp, pp)
## for mu
XX[i1,] <- matrix(c(1,  1,  0), nn, pp, byrow = TRUE)
## for beta_1
XX[i2,] <- matrix(c(1,  0,  1), nn, pp, byrow = TRUE)
## for beta_2
XX[i3,] <- matrix(c(1, -1, -1), nn, pp, byrow = TRUE)
## fit model & get parameters
Bvec <- coef(lm(yy ~ XX - 1))
names(Bvec) <- c("mu", "beta_1", "beta_2")
Bvec
```

```
##        mu    beta_1    beta_2
## 26.761236 -7.141222 -1.112776
```

Defining models with $\mathbf{X}$

ANOVA

```
## mean of ration 1
Bvec["mu"] + Bvec["beta_1"]
## mean of ration 2
Bvec["mu"] + Bvec["beta_2"]
## mean of ration 3
Bvec["mu"] - (Bvec["beta_1"] + Bvec["beta_2"])
```

```
##       mu
## 19.62001
##       mu
## 25.64846
##       mu
## 35.01523
```

Defining models with $\mathbf{X}$

ANOVA

We could also fit our grand mean model after some simple algebra

$$
\begin{gathered}
y_i = \mu + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i \\
\Downarrow \\
y_i - \mu = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i \\
\Downarrow \\
y_i - \bar{y} = \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + e_i
\end{gathered}
$$

Defining models with $\mathbf{X}$

ANOVA

```
## fit anova with implicit grand mean
m2 <- lm((yy - mean(yy)) ~ ration - 1)
coef(m2)
```

```
##  ration_1  ration_2  ration_3 
## -7.141222 -1.112776  8.253998
```

Defining models with $\mathbf{X}$

ANOVA

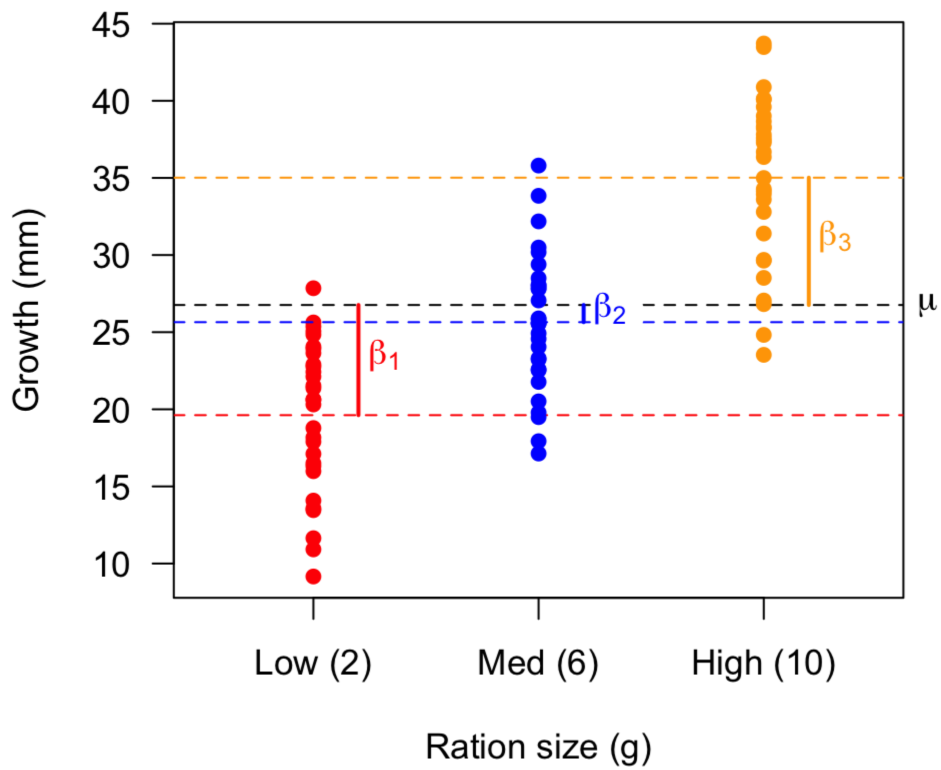
```r
## do we recover our means?
coef(m2) + mean(yy)
```

```
## ration_1 ration_2 ration_3
## 19.62001 25.64846 35.01523
```

```r
coef(m1)
```

```
## ration_1 ration_2 ration_3
## 19.62001 25.64846 35.01523
```

Comparing group means



Defining models with $\mathbf{X}$

ANOVA

What if we wanted to treat one group as a control or reference (eg, our low ration) and estimate the other effects relative to it?

$$y_i = \beta_1 x_{1,i} + (\beta_1 + \beta_2) x_{2,i} + (\beta_1 + \beta_3) x_{3,i} + e_i$$

such that

$$\begin{aligned}
\bar{y}_{j=1} &= \beta_1 \\
\bar{y}_{j=2} &= \beta_1 + \beta_2 \\
\bar{y}_{j=3} &= \beta_1 + \beta_3
\end{aligned}$$

Defining models with $\mathbf{X}$

We would define $\mathbf{X}$ and $\boldsymbol{\beta}$ as

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ \hline 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ \hline 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

Defining models with $\mathbf{X}$

ANOVA

```r
## empty design matrix
XX <- matrix(NA, nn*pp, pp)
## for beta_1
XX[i1,] <- matrix(c(1, 0, 0), nn, pp, byrow = TRUE)
## for beta_1 + beta_2
XX[i2,] <- matrix(c(1, 1, 0), nn, pp, byrow = TRUE)
## for beta_1 + beta_3
XX[i3,] <- matrix(c(1, 0, 1), nn, pp, byrow = TRUE)
## fit anova with implicit grand mean
Bvec <- coef(lm(yy ~ XX - 1))
names(Bvec) <- c("beta_1", "beta_2", "beta_3")
Bvec
```

```
##    beta_1    beta_2    beta_3
## 19.620014  6.028446 15.395221
```

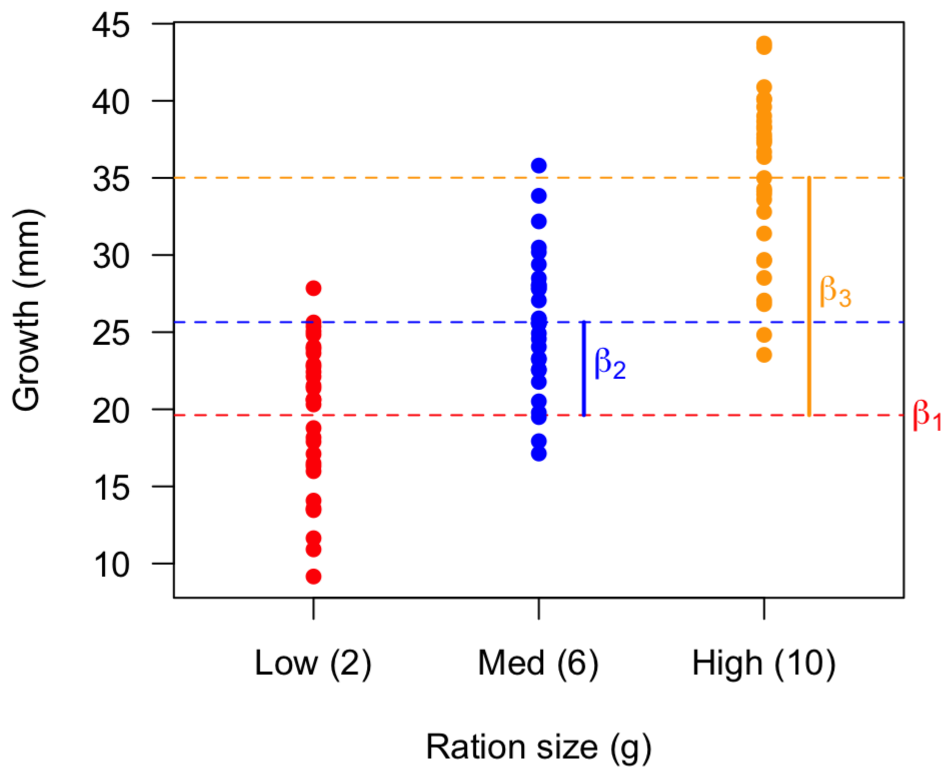
Defining models with $\mathbf{X}$

ANOVA

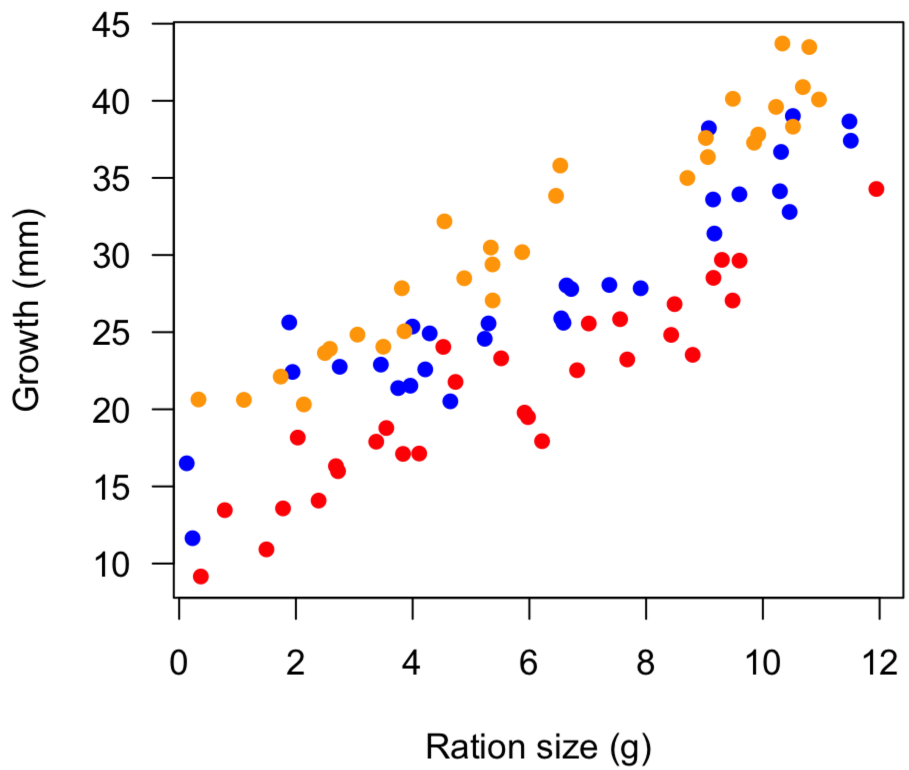
```
## mean of ration 1
Bvec["beta_1"]
## mean of ration 2
Bvec["beta_1"] + Bvec["beta_2"]
## mean of ration 3
Bvec["beta_1"] + Bvec["beta_3"]
```

```
##   beta_1
## 19.62001
##   beta_1
## 25.64846
##   beta_1
## 35.01523
```

Comparing group means



Analysis of covariance (ANCOVA)



Analysis of covariance (ANCOVA)

Here is our model with the categorical effect of lineage & the continuous effect of ration

$$\text{growth}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}_{1,\text{lineage}} + \boldsymbol{\beta}_2 \text{ration}_i + \boldsymbol{\epsilon}_i$$

Analysis of covariance (ANCOVA)

Dropping the global intercept & writing out the lineage effects yields

$$\text{growth}_{\boldsymbol{i}} = \underbrace{\boldsymbol{\beta}_1 \boldsymbol{x}_{1,\boldsymbol{i}} + \boldsymbol{\beta}_2 \boldsymbol{x}_{2,\boldsymbol{i}} + \boldsymbol{\beta}_3 \boldsymbol{x}_{3,\boldsymbol{i}}}_{\text{lineage}} + \underbrace{\boldsymbol{\beta}_4 \boldsymbol{x}_{4,\boldsymbol{i}}}_{\text{ration}} + \boldsymbol{e}_{\boldsymbol{i}}$$

Defining models with $\mathbf{X}$

We would then define $\mathbf{X}$ and $\boldsymbol{\beta}$ as

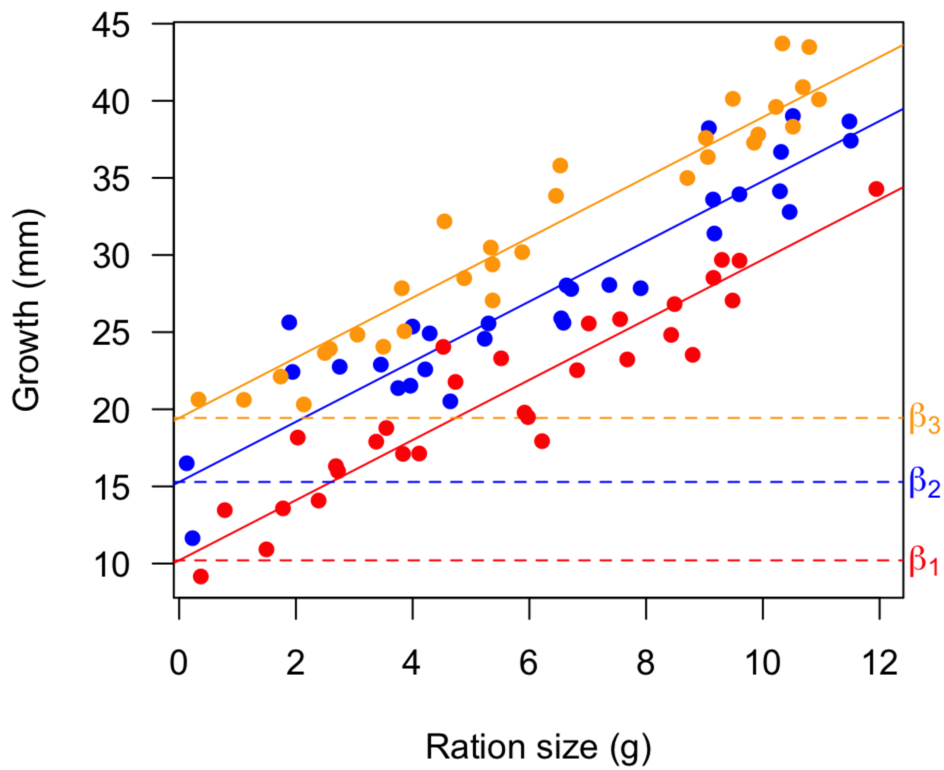
$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & \boldsymbol{r}_1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \boldsymbol{r}_{j_1} \\ \hline 0 & 1 & 0 & \boldsymbol{r}_{j_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \boldsymbol{r}_{j_2+j_2} \\ \hline 0 & 0 & 1 & \boldsymbol{r}_{j_1+j_2+1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & \boldsymbol{r}_n \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

Analysis of covariance (ANCOVA)

```r
## create design matrix
XX <- cbind(L1 = rep(c(1,0,0), ea = nn), # effect of lineage 1
            L2 = rep(c(0,1,0), ea = nn), # effect of lineage 2
            L3 = rep(c(0,0,1), ea = nn), # effect of lineage 3
            RA = x_cov)                  # effect of ration
## fit model
Bvec <- coef(lm(yy ~ XX - 1))
names(Bvec) <- c("beta_1", "beta_2", "beta_3", "beta_4")
Bvec
```

```
##    beta_1    beta_2    beta_3    beta_4
## 10.205959 15.286507 19.435551  1.950062
```

Analysis of covariance (ANCOVA)



Design matrices with `model.matrix()`

We have been building our design matrices by hand, but we could instead use

`model.matrix()` with `factor()`

Design matrices with `model.matrix()`

`factor(x)` tells **R** to treat `x` as categorical

```
## 2 groups with 2 obs each
groups <- factor(c(1, 1, 2, 2))
## inspect them
groups
```

```
## [1] 1 1 2 2
## Levels: 1 2
```

Design matrices with `model.matrix()`

`model.matrix(~ x)` uses a right-hand side formula `~ x`

```
## create design matrix from `groups`
model.matrix(~ groups)
```

```
##   (Intercept) groups2
## 1           1       0
## 2           1       0
## 3           1       1
## 4           1       1
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$groups
## [1] "contr.treatment"
```

Design matrices with `model.matrix()`

What if we don't use `factor()`?

```r
## 2 groups with 2 obs each
groups <- c(1, 1, 2, 2)
## create design matrix from `groups`
model.matrix(~ groups)
```

```
##   (Intercept) groups
## 1           1      1
## 2           1      1
## 3           1      2
## 4           1      2
## attr(,"assign")
## [1] 0 1
```

Design matrices with `model.matrix()`

You can drop the intercept term with `- 1`

```r
## 2 groups with 2 obs each
groups <- factor(c(1, 1, 2, 2))
## create design matrix from `groups`
model.matrix(~ groups - 1)
```

```
##   groups1 groups2
## 1       1       0
## 2       1       0
## 3       0       1
## 4       0       1
## attr(,"assign")
## [1] 1 1
## attr(,"contrasts")
## attr(,"contrasts")$groups
## [1] "contr.treatment"
```

Design matrices with `model.matrix()`

The names/categories are irrelevant for `factor()`

```r
## 2 groups with 2 obs each
groups <- factor(c("ref", "ref", "exp", "exp"))
## create design matrix from `groups`
model.matrix(~ groups)
```

```
##   (Intercept) groupsref
## 1           1         1
## 2           1         1
## 3           1         0
## 4           1         0
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$groups
## [1] "contr.treatment"
```

Design matrices with `model.matrix()`

R assigns factors in alphabetical order; the *reference* is first

```
## 2 groups with 2 obs each
groups <- factor(c("ref", "ref", "exp", "exp"))
## create design matrix from `groups`
model.matrix(~ groups)
```

```
##   (Intercept) groupsref
## 1           1         1
## 2           1         1
## 3           1         0
## 4           1         0
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$groups
## [1] "contr.treatment"
```

Design matrices with `model.matrix()`

We can change the reference case with `relevel()`

```
## 2 groups with 2 obs each
groups <- relevel(groups, "ref")
## create design matrix from `groups`
model.matrix(~ groups)
```

```
##   (Intercept) groupsexp
## 1           1         0
## 2           1         0
## 3           1         1
## 4           1         1
## attr(,"assign")
## [1] 0 1
## attr(,"contrasts")
## attr(,"contrasts")$groups
## [1] "contr.treatment"
```

Design matrices with `model.matrix()`

We can add multiple factors with `+`

```
diet <- factor(c(1, 1, 2, 2))
sex <- factor(c("f", "m", "f", "m"))
model.matrix(~ diet + sex)
```

```
##   (Intercept) diet2 sexm
## 1           1     0    0
## 2           1     0    1
## 3           1     1    0
## 4           1     1    1
## attr(,"assign")
## [1] 0 1 2
## attr(,"contrasts")
## attr(,"contrasts")$diet
## [1] "contr.treatment"
##
## attr(,"contrasts")$sex
## [1] "contr.treatment"
```

Design matrices with `model.matrix()`

You can also extract the design matrix from a fitted model

```
## ANCOVA model from above
mod_fit <- lm(yy ~ XX - 1)
## get design matrix
mm <- model.matrix(mod_fit)
head(mm)
```

```
##   XXL1 XXL2 XXL3      XXRA
## 1    1    0    0 11.944444
## 2    1    0    0  3.835147
## 3    1    0    0  3.376075
## 4    1    0    0  4.112188
## 5    1    0    0  2.721664
## 6    1    0    0  1.779256
```