Lesson 3

Mixed effects models:

An introduction

Assumptions of linear models

$$y_{ijk} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + e_{ijk}$$

Linear models make many assumptions, including:

1. The model makes biological sense/ physical sense
2. Additivity (terms are added together)
3. Linearity
4. Independence of errors
5. Homoscedasticity – equal variance of errors
6. Normality of errors.

Fixed and random effects

- Until now, we have treated all predictor variables as equal.
- However, we can distinguish two types of fundamentally different predictor variables called:
- Fixed effects,   and
- Random effects

Fixed effects vs Random effects

Fixed effects

- are the predictor variables that we are interested in.
- We want to understand their effect on the response variable (y).
- They are specifically referred to in our hypotheses.
- We can quantify/qualify their values.

- e.g.
- Q: Do forest types differ in their understory productivity?
- Predictor: 3 Forest types: beech, podocarp, mixed
- Response: understory biomass

Fixed Effects

- Are we interested in effect sizes?
- Are the factor levels informative?
- Are the factor levels experimental manipulations?

- If yes, it is a fixed effect

Fixed Effects vs Random effects

<u>Random effects</u>

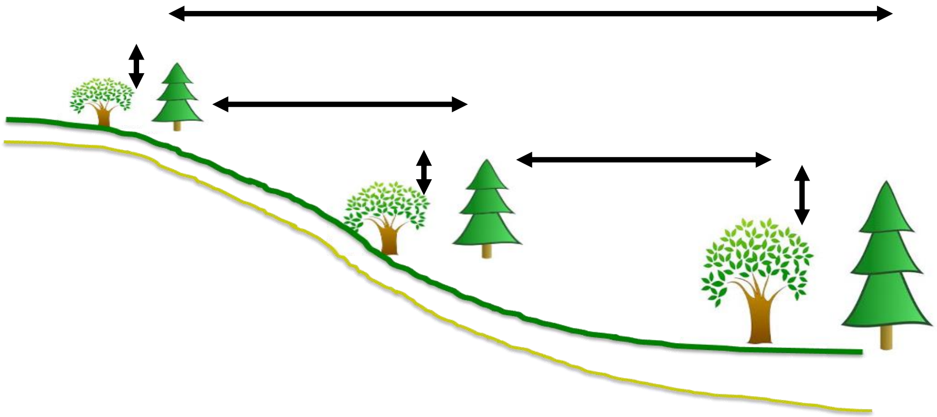
- are additional variables we include in the model which can explain some of the variance in Y.
- We are not interested in them, but they are useful for two reasons:
  - They <u>increase our test power</u> as they <u>reduce the residual error</u> by explaining some variance in our data
  - More importantly, they correct for dependence between data points

Why use MMs?

- To properly account for dependence structure of data and thus sidestep pseudoreplication (dependency)
- To treat fixed and random effects appropriately
  - Fixed effects: Estimate and test
  - Random effects: Predict and test variance components
- Get more appropriate estimates (including benefits of shrinkage -> next lecture)
- Get better residual distributions
- To circumvent missing-value problems associated with ANOVA

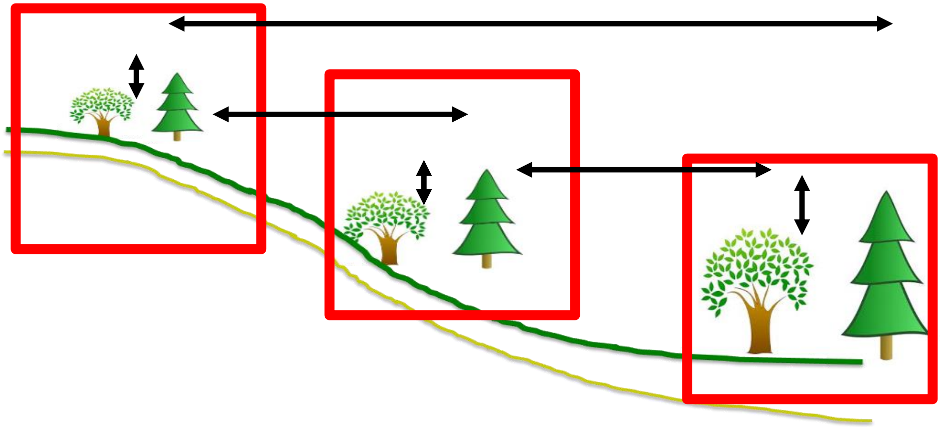
Example: Spatial blocks

- Height differences between tree species



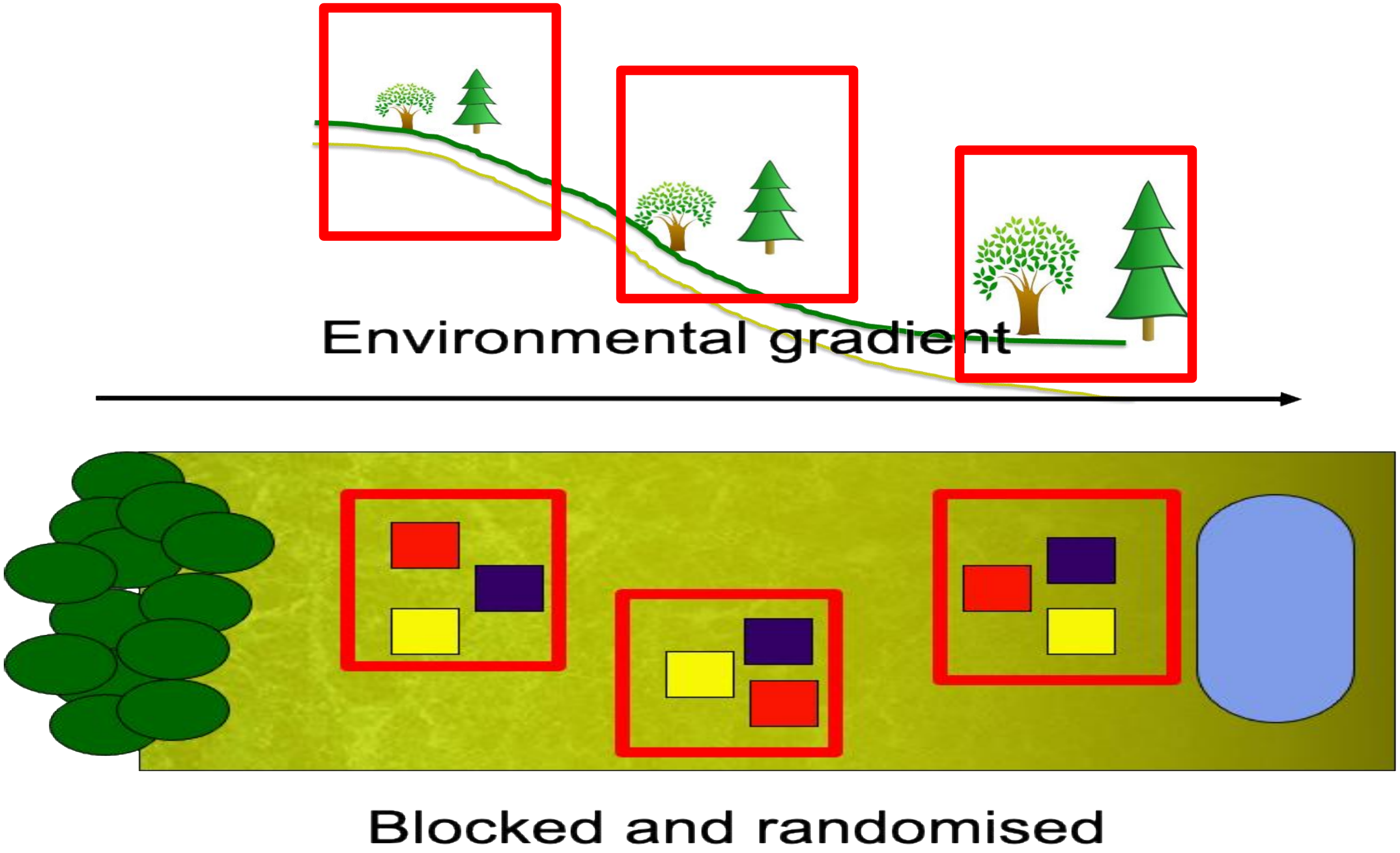
Example: Spatial blocks

- Height differences between tree species



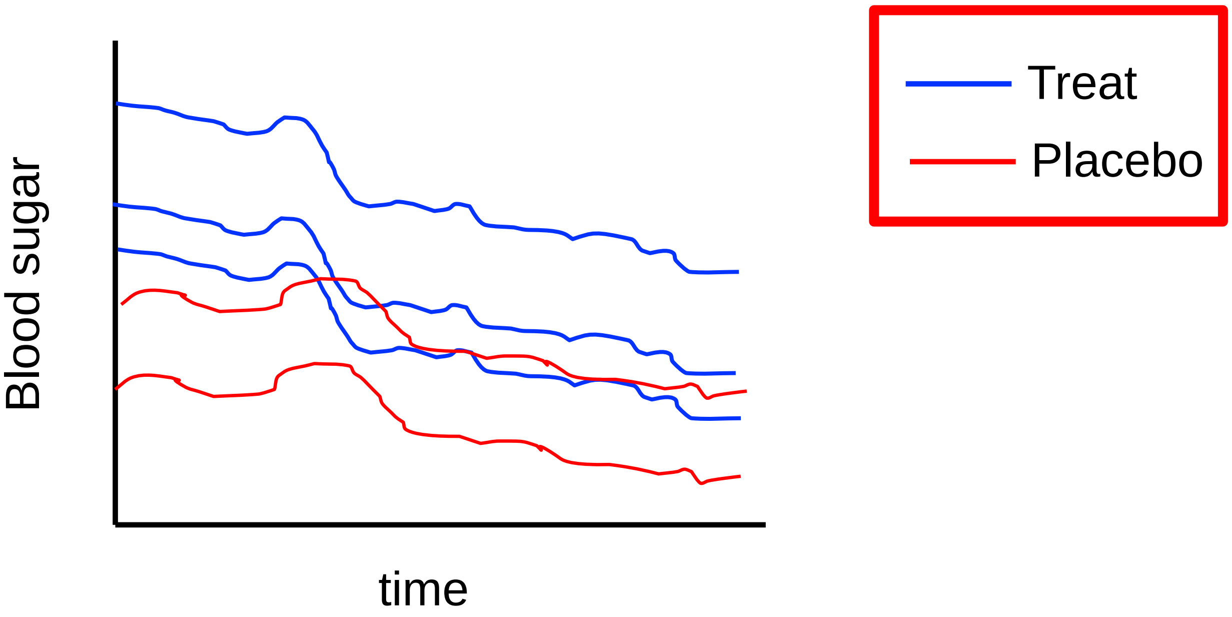
Example: Spatial blocks

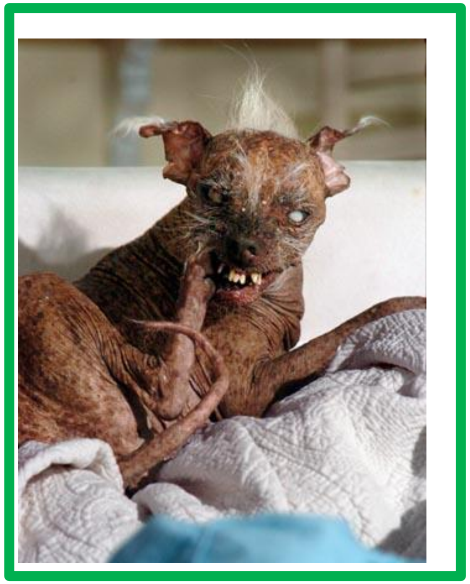
- Height differences between tree species



Example: subject effects

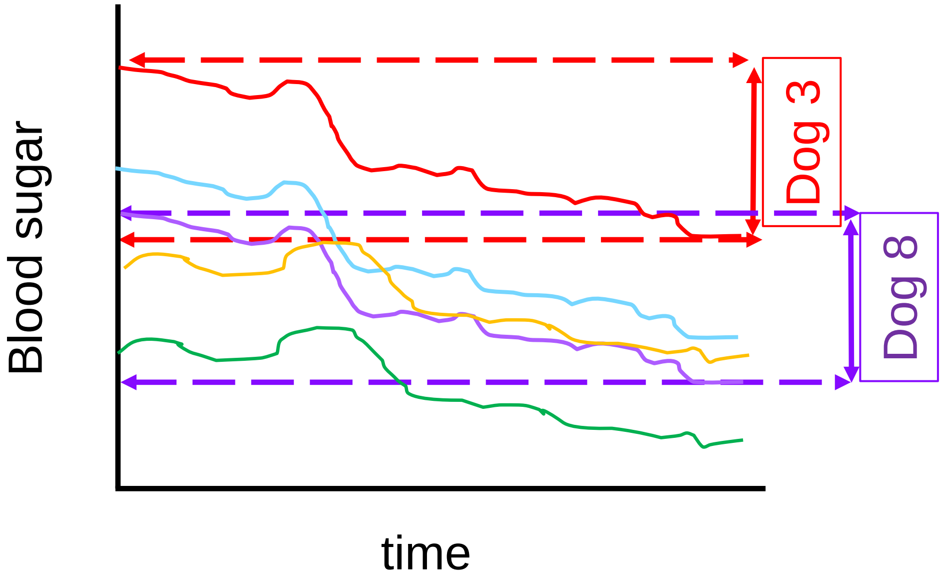
- Experiment looking at drug effect on blood sugar over time in dogs
- Each individual dog gets a specific treatment (the treatment is the fixed effect we are interested in)





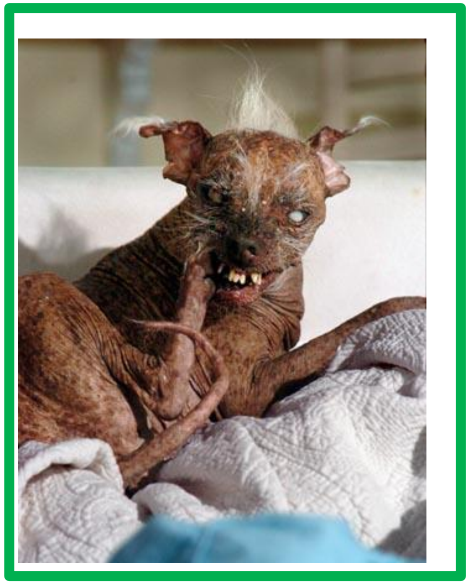
Example: subject effects

- Experiment looking at drug effect on blood sugar <u>over time </u>in dogs
- Each individual has a unique trait range (data pts from each dog are linked -> random effect)



MM: fixed and random effects

- Fixed effect: predictors we care about
  - E.g. drug treatment, time (in our Hypotheses/ questions)
- Random effect: predictors we don’t care about, but which might be causing variation in the data e.g. dogID



Linear MMs in R: lmer() statement

There are numerous packages in R for mixed model analysis , but we will mostly consider lme4 here.

Consider a crop yield experiment with two fertilizer treatments, each with multiple levels (N, P) and fully crossed in repeated blocks.

The lmer() model statement specifies the fixed factors (N*P) and the structure of the random factors (1|Block).

- M2 <- lmer(Yield ~ N*P + (1|Block),df3)

Mixed models: lmer() output

```
> summary(M2)
Linear mixed model fit by REML
Formula: Yield ~ N * P + (1 | Block)
Data: df3
   AIC   BIC logLik deviance REMLdev
 264.3 271.3 -127.1    270.4   254.3
Random effects:
 Groups   Name        Variance Std.Dev.
 Block    (Intercept)  90.899   9.5341
 Residual             470.305  21.6865
Number of obs: 30, groups: Block, 5

Fixed effects:
            Estimate Std. Error t value
(Intercept)    7.500     16.401   0.457
N             22.267      7.919   2.812
P              0.650      4.849   0.134
N:P           -8.700      9.741  -0.893
```

For random effects, only their variance components are shown. Why? Because random effects are assumed to have mean = 0 (or rather that their means are unimportant to us).

Mixed models deal with unbalanced designs (among other problems), for which error df need to be estimated. Therefore the lme4 package does not provide probabilities automatically.

LMMs:
very simplified mathematics

- Mixed models are linear models with a component for fixed factors and variates (Xβ) and a component for random factors (Zb)

$$\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\varepsilon}, \qquad \mathbf{b} \sim N(0,\Psi), \qquad \boldsymbol{\varepsilon} \sim N(0,\sigma^2)$$

- Importantly, the coefficients of the random variables are assumed to have a normal distribution around a mean of zero.
- We are only interested in removing the variance (Ψ) the random effect levels account for.

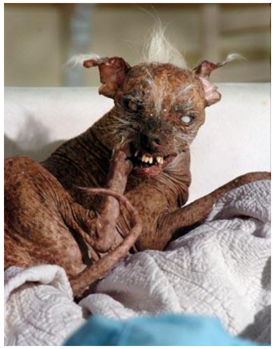
$y = X\beta + Zb + \varepsilon$ expanded

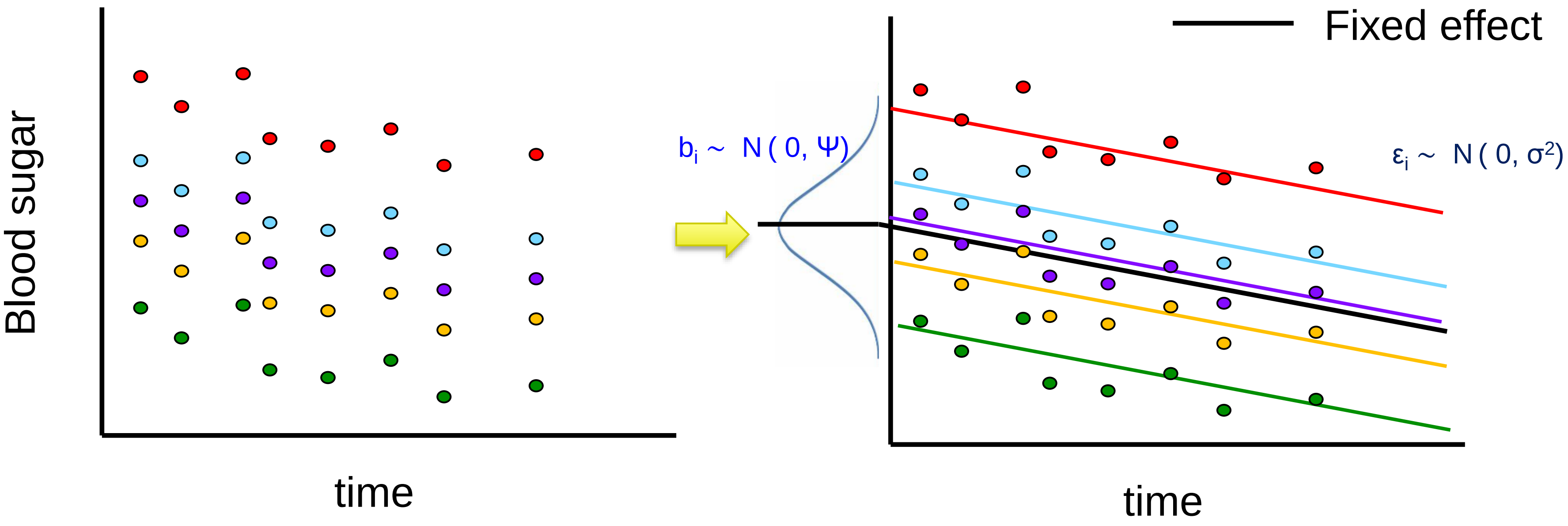
$$\mathbf{Y} \quad = \quad \mathbf{X} \quad * \quad \boldsymbol{\beta} \quad + \quad \mathbf{Z} \quad * \quad \mathbf{b} \quad + \quad \boldsymbol{\varepsilon}$$

$$\underset{(nx1)}{\begin{bmatrix} y_1\\ y_2\\ y_3\\ .\\ .\\ y_{n-1}\\ y_n \end{bmatrix}} = \underset{(nxp)}{\begin{array}{c} \begin{array}{ccc} \text{Int} & \text{X1} & \text{X2} \end{array}\\ \begin{bmatrix} 1 & x_{1,1} & x_{2,1}\\ 1 & x_{1,2} & x_{2,2}\\ 1 & x_{1,3} & x_{2,3}\\ & . & .\\ & . & .\\ 1 & x_{1,n-1} & x_{2,n-1}\\ 1 & x_{1,n} & x_{2,n} \end{bmatrix} \end{array}} \underset{(px1)}{\begin{bmatrix} \beta_0\\ \beta_1\\ \beta_2 \end{bmatrix}} + \underset{(nxq)}{\begin{array}{c} \begin{array}{cccc} \text{Z1} & \text{Z2} & \text{Z3} & \text{Z4} \end{array}\\ \begin{bmatrix} 1 & 0 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}} \underbrace{\underset{(qx1)}{\begin{bmatrix} b_1\\ b_2\\ b_3\\ b_4 \end{bmatrix}}}_{\mathbf{b} \sim N(0,\Psi)} + \underbrace{\underset{(nx1)}{\begin{bmatrix} \varepsilon_1\\ \varepsilon_2\\ \varepsilon_3\\ .\\ .\\ \varepsilon_{n-1}\\ \varepsilon_n \end{bmatrix}}}_{\boldsymbol{\varepsilon} \sim N(0,\sigma^2)}$$

MM: geometry

- Response (Y): Blood sugar
- Fixed effect: time
- Random effect: dogID
- Model: Sugar ~ time + (1|dogID)





MMs are solved with ML/REML

- Mixed models are solved with Restricted maximum likelihood (REML) to get unbiased estimates of variances, and maximum likelihood (ML) for nested model comparison (LRTs, AIC, AICc)

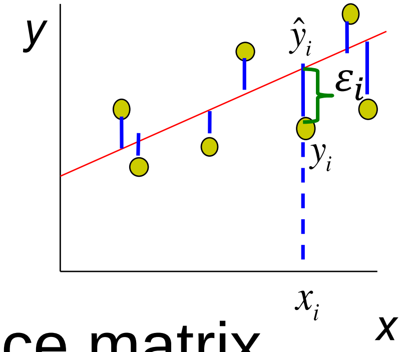
- I do not fully understand the maths behind REML so i will explain the ML estimation here as ML and REML yield identical estimates for the coefficients (β_i)

Recall MLE for the linear model

- The linear model has the following form

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{where} \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 I)$$

- Remember that **ε** is a vector of errors around the fitted line.



- now $E(\mathrm{var}(\boldsymbol{\varepsilon})) = \sigma^2$, $V = \sigma^2 I$, the data covariance matrix

$$\begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \varepsilon_3 & . & . & \varepsilon_{n-1} & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ \varepsilon_{n-1} \\ \varepsilon_n \end{bmatrix} \mathrm{x} \begin{bmatrix} 1 & 0 & 0 & \dots & 0 & 0 \\ 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ . & . & . & & . & . \\ . & . & . & & . & . \\ 0 & 0 & 0 & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix} = \begin{bmatrix} \varepsilon_{1,1}^2 & 0 & 0 & \dots & 0 & 0 \\ 0 & \varepsilon_{2,2}^2 & 0 & \dots & 0 & 0 \\ 0 & 0 & \varepsilon_{3,3}^2 & \dots & 0 & 0 \\ . & . & . & & . & . \\ . & . & . & & . & . \\ 0 & 0 & 0 & \dots & \varepsilon_{n-1.n-}^2 & 0 \\ 0 & 0 & 0 & \dots & 0 & \varepsilon_{n,n}^2 \end{bmatrix}$$

Recall MLE for the linear model

- The simple linear model has a pdf of the form

$$f(\varepsilon_i) = \frac{1}{\sigma\sqrt{2\pi}}\, exp\left(-\frac{1}{2\sigma^2}\varepsilon_i{}^2\right)$$

- This converts to a likelihood of the form for n observations

$$L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = \frac{1}{(2\pi)^{n/2}\sigma^n}\, exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'(\sigma^2 I)^{-\mathbf{1}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)$$

- Now as $E(var(\boldsymbol{\varepsilon}))=\sigma^2$, so $\sigma^2 I = V$, and

$$L(\boldsymbol{y},\boldsymbol{X},\boldsymbol{\beta},\sigma^2) = \frac{1}{(2\pi)^{n/2}V^{n/2}}\, exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-\mathbf{1}}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right)$$

- Solving this yields coefficient estimates:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{y} = (\boldsymbol{X}'\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}'\boldsymbol{y} \quad [\boldsymbol{because}\ \boldsymbol{V} = \sigma^2 I\ ]$$

MLE for the MM model

- In the MM case, we have the additional term Zb

  $\mathbf{y} = X\boldsymbol{\beta} + Z\mathbf{b} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(0, \sigma_r^2 I)$, $\mathbf{b} \sim N(0, \sigma_q^2)$

- Now the covariance matrix is the form: $V = Z\sigma_q^2 + I\sigma_r^2$

- So the likelihood takes the same basic form

$$L\left(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma_q^2, \sigma_r^2\right) = \frac{1}{(2\pi)^{n/2} V^{n/2}}\, exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})' \boldsymbol{V}^{-\mathbf{1}} (\boldsymbol{y} - \boldsymbol{X\beta})\right)$$

- Which yields the same general solution for coefficients:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{X}\right)^{-\mathbf{1}} \boldsymbol{X}'\boldsymbol{V}^{-\mathbf{1}}\boldsymbol{y}$$

- (V is more complicated than I write it here)

End of Lesson 3