NTNU | Norwegian University of Science and Technology

REGRESSION DIAGNOSTICS

from regression to GLMMs and beyond

Bert van der Veen

Outline

Residuals diagnostics in

1. LMs
2. GLMs
3. GLMMs
4. GLLVMs

Slides: https://github.com/BertvanderVeen/Educational-material

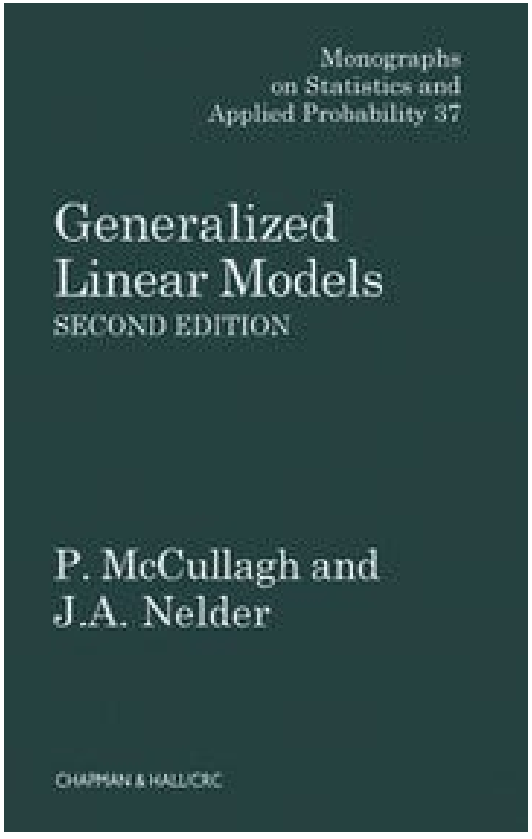
Things that I will ignore

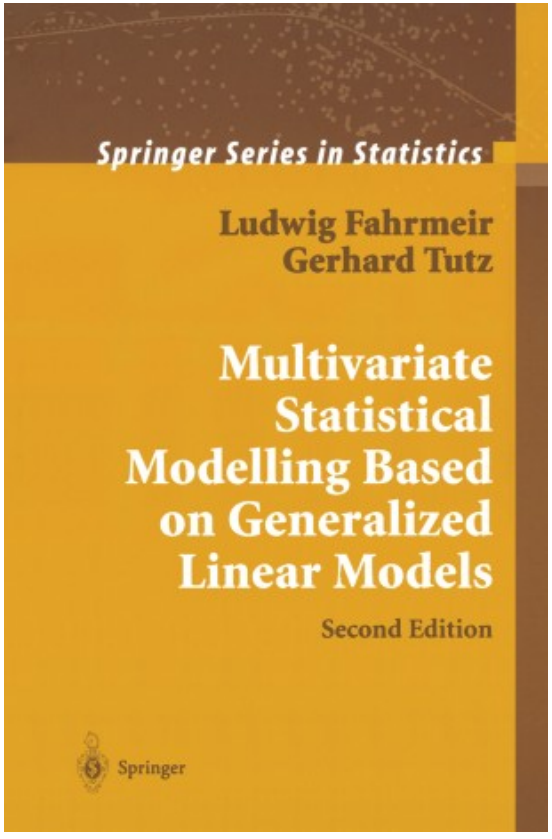
Other diagnostics tools

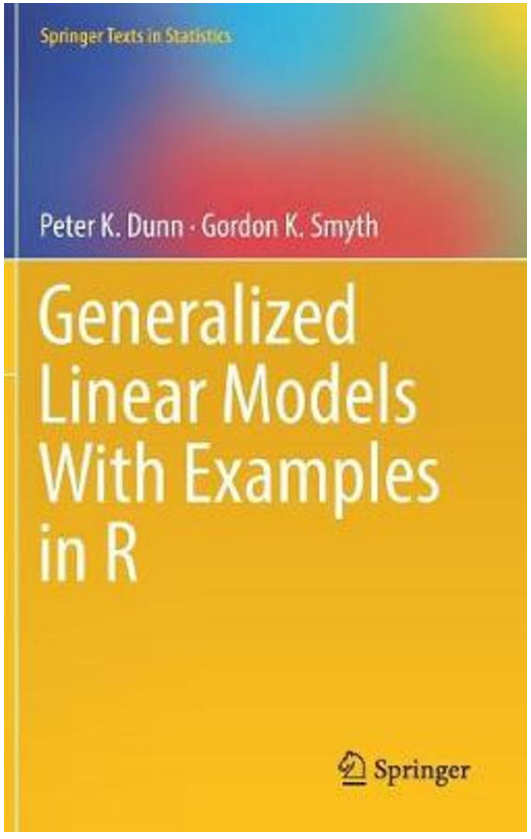
- R^2
- Variance Inflation Factor (VIF)
- Cross validation
- Nested and non nested hypotheses tests
- Statistical uncertainties
- Information criteria (AIC, BIC)

Ecological relevance of assumptions/violations

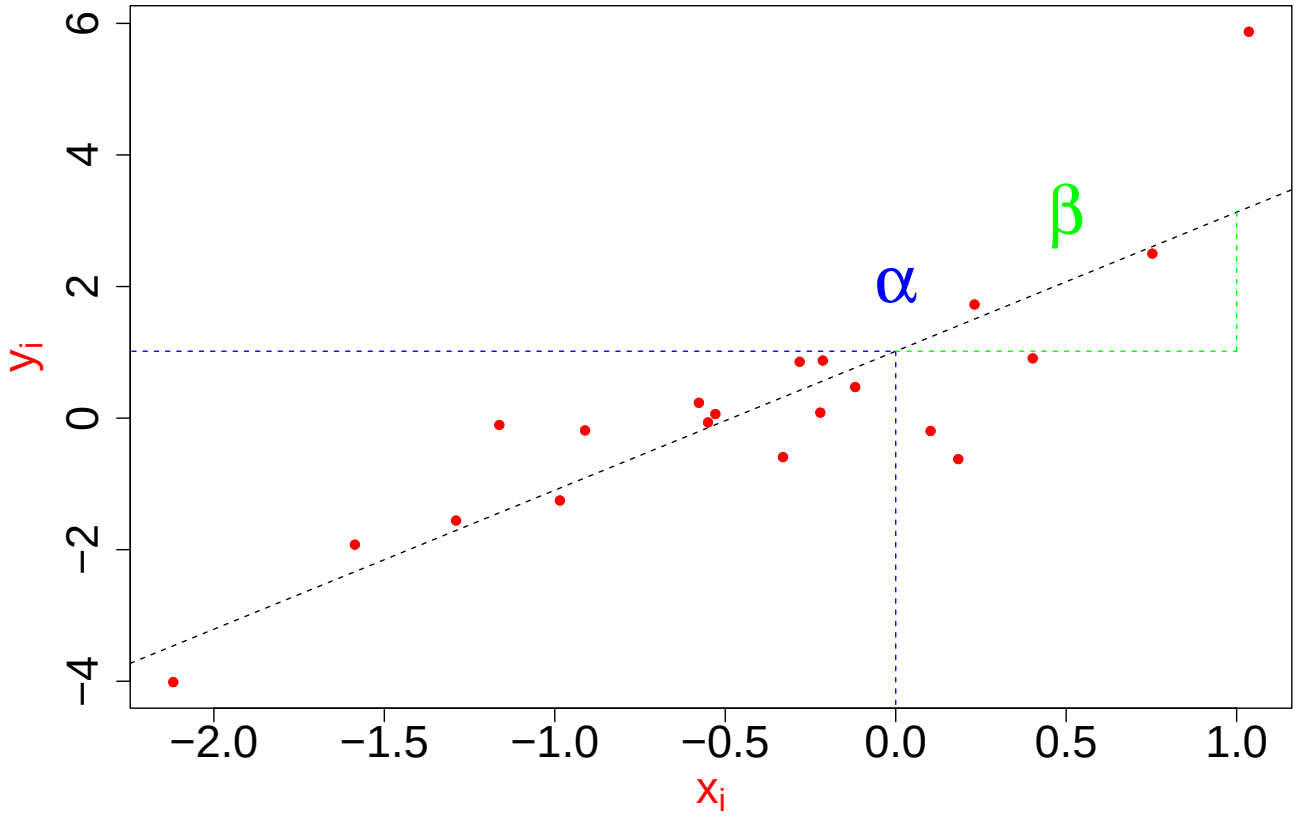
Some resources



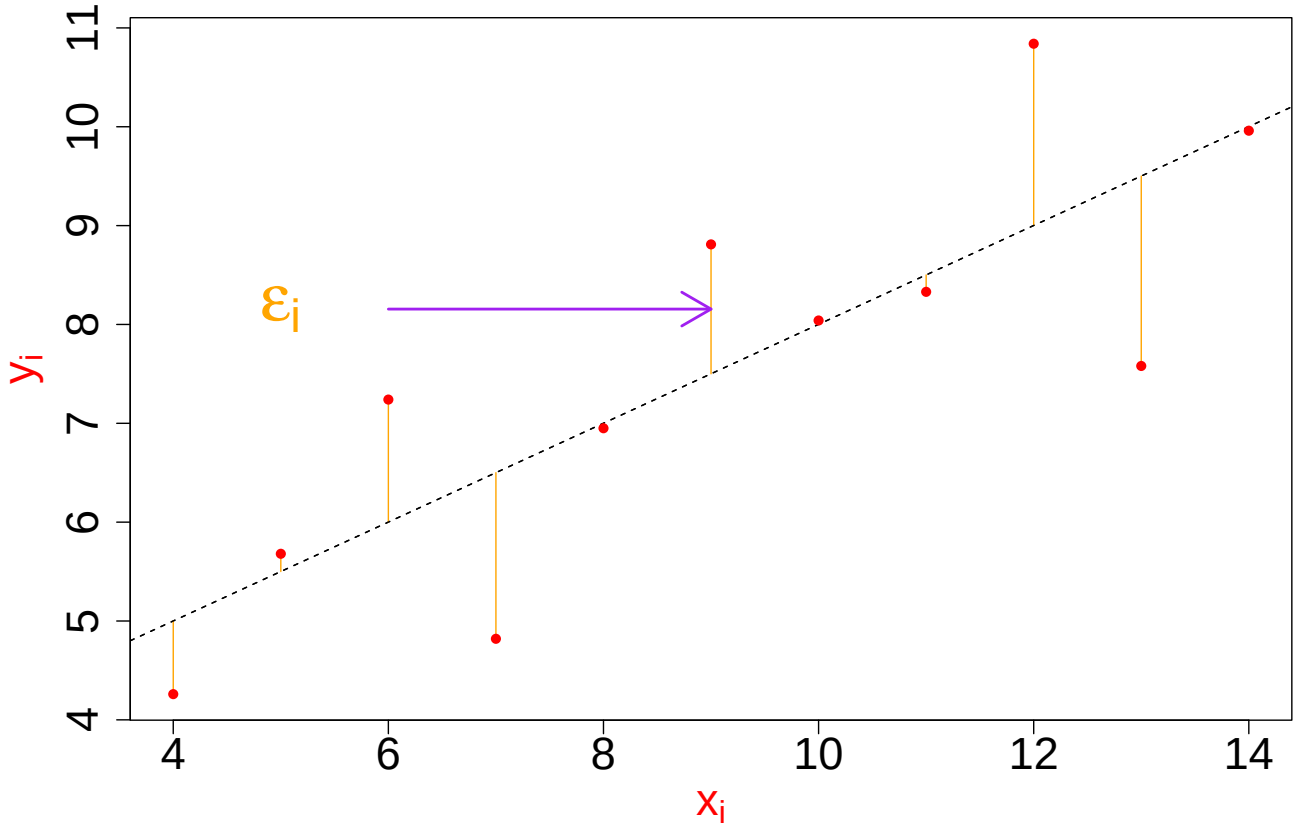




Quick recap: α and β



Quick recap: the error ϵ_i



Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

i.e. $\mu = \mathsf{E}(y_i|x_i) = \alpha + \beta x_i$

Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

i.e. $\mu = \mathsf{E}(y_i|x_i) = \alpha + \beta x_i$

1. Data y_i, x_i

Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

i.e. $\mu = \mathsf{E}(y_i|x_i) = \alpha + \beta x_i$

1. Data y_i, x_i
2. Intercept α

Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

i.e. $\mu = \mathsf{E}(y_i|x_i) = \alpha + \beta x_i$

1. Data y_i, x_i
2. Intercept α
3. Slope β

Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

i.e. $\mu = \mathsf{E}(y_i|x_i) = \alpha + \beta x_i$

1. Data y_i, x_i
2. Intercept α
3. Slope β
4. Error ϵ_i

Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

i.e. $\mu = \mathsf{E}(y_i|x_i) = \alpha + \beta x_i$

1. Data y_i, x_i
2. Intercept α
3. Slope β
4. Error ϵ_i
   - 4.1 Error variance σ^2

Quick recap: linear regression

$$y_i = \alpha + \beta x_i + \epsilon_i \sim \mathcal{N}(0, \sigma^2) \equiv y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2) \qquad (1)$$

i.e. $\mu = \mathsf{E}(y_i|x_i) = \alpha + \beta x_i$

1. Data y_i, x_i
2. Intercept α
3. Slope β
4. Error ϵ_i
   - 4.1 Error variance σ^2

Assumptions

- Linearity
- Normality, constant variance
- Independence of errors
- Absence of (perfect) multicollinearity
- No error in x_i
- Lack of outliers

Quick recap: estimating α and β

- Found by minimizing $RSS = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$ in LM

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i \tag{2}$$

Quick recap: estimating α and β

- Found by minimizing $RSS = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$ in LM

$$\hat{\mu}_i = \hat{\alpha} + \hat{\beta} x_i \tag{2}$$

  - $\hat{\beta} = (X^\top X)^{-1} X^\top y$, the BLUE
    - unbiased and minimum variance
    - estimators exist with less variance (e.g. ridge)
  - $\hat{\beta} \sim \mathcal{N}(\beta, (X^\top X)^{-1} \sigma^2)$
    - without constant variance biased SE

Linear regression: $\widehat{\epsilon}_i$

$$\widehat{\epsilon}_i = y_i - \widehat{\mu}_i \tag{3}$$

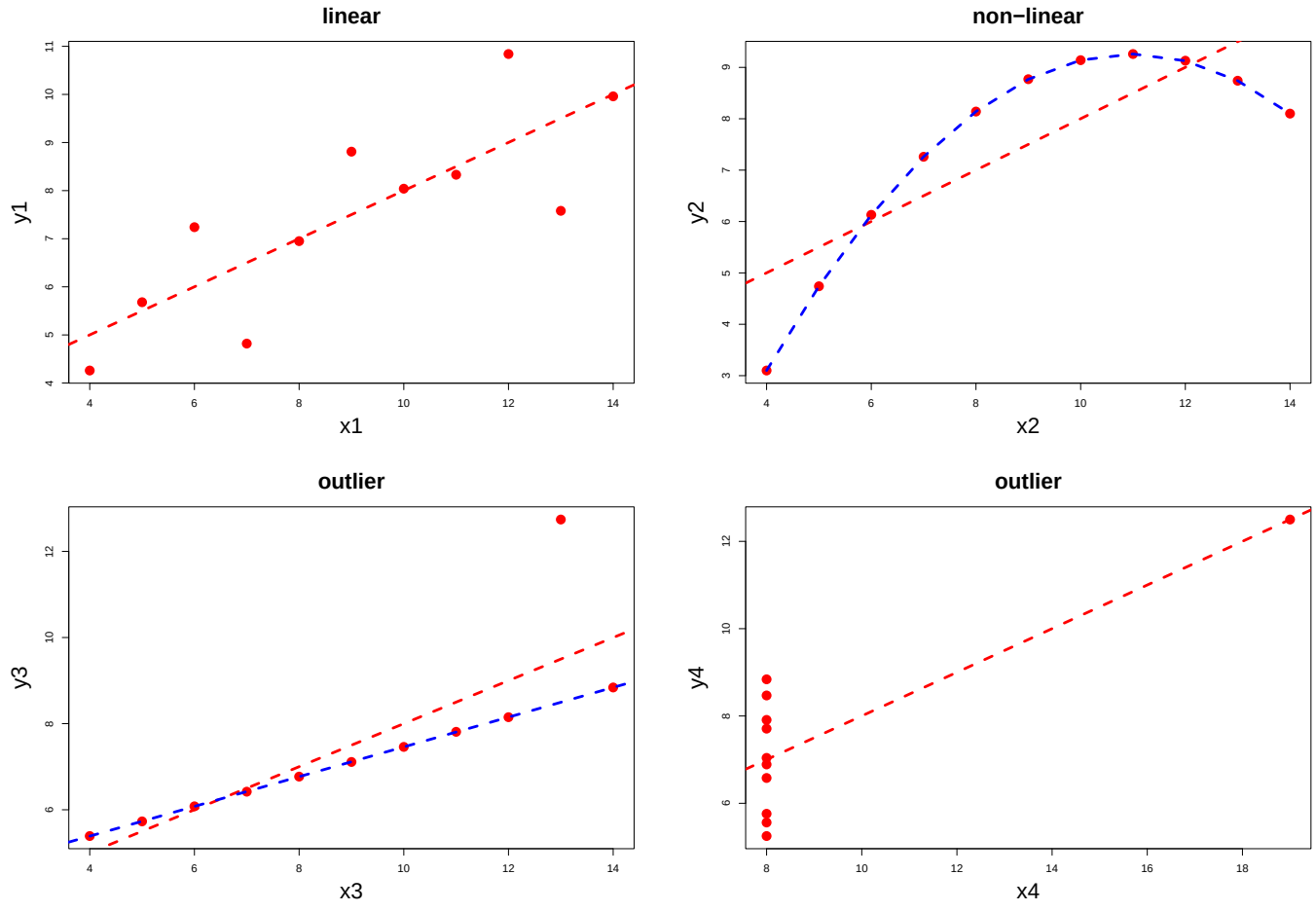
Standardised residuals

$$\frac{\widehat{\epsilon}_i}{s\sqrt{1-h_i}}, \qquad \text{where } h_i \text{ is "influence"} \tag{4}$$

Studentized residuals

- omit observation i
- re-compute residual variance
- calculate standardised residual (with hat matrix)

Anscombe quartet



Anscombe quartet: everything equal!

```
all.equal(mean(anscombe$x1),mean(anscombe$x2),mean(anscombe$x3),mean(anscombe$x4))
```

```
## [1] TRUE
```

```
all.equal(mean(anscombe$y1),mean(anscombe$y2),mean(anscombe$y3),mean(anscombe$y4))
```

```
## [1] TRUE
```

```
all.equal(sd(anscombe$x1),sd(anscombe$x2),sd(anscombe$x3),sd(anscombe$x4))
```

```
## [1] TRUE
```

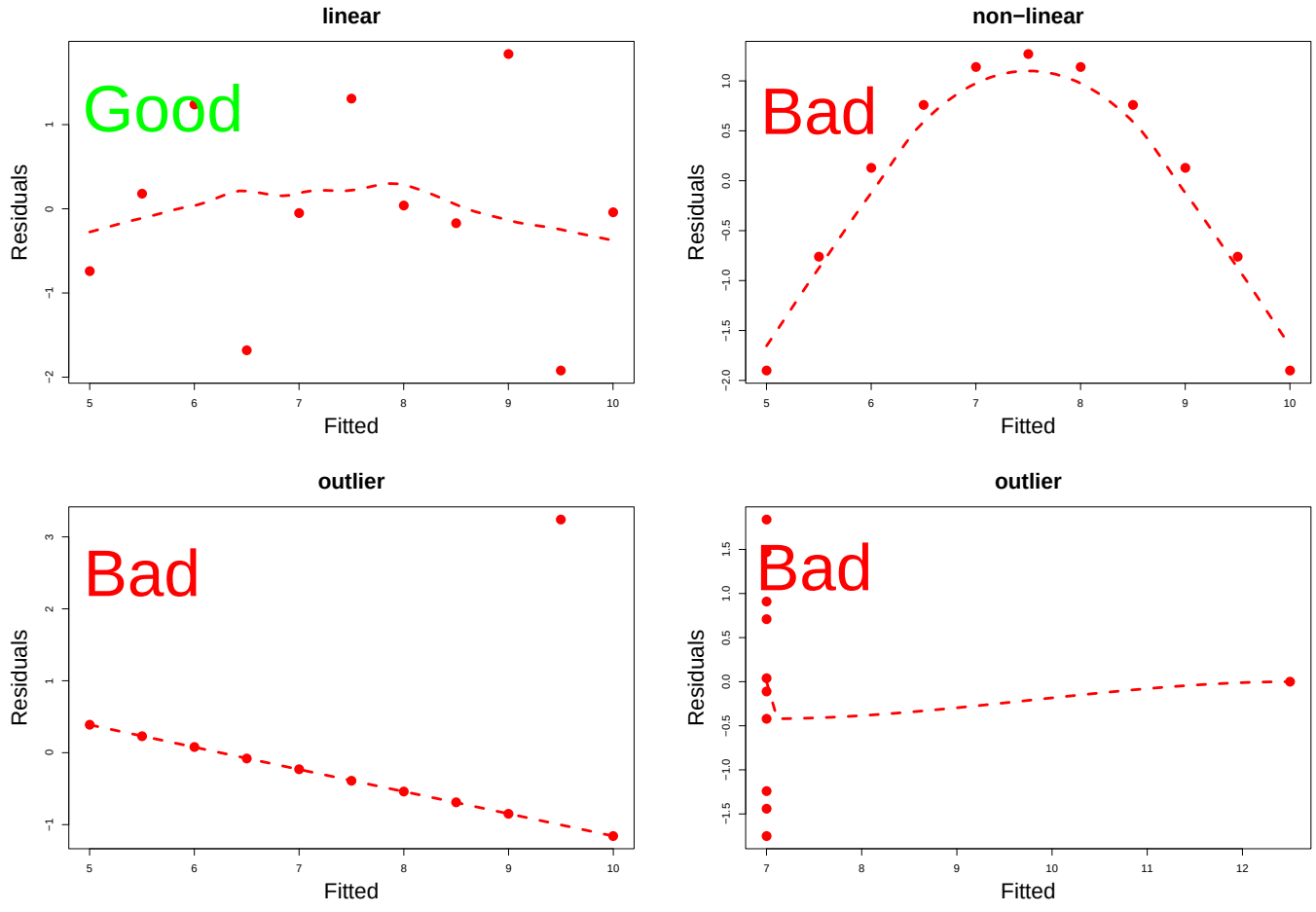
```
all.equal(sd(anscombe$y1),sd(anscombe$y2),sd(anscombe$y3),sd(anscombe$y4))
```

```
## [1] TRUE
```

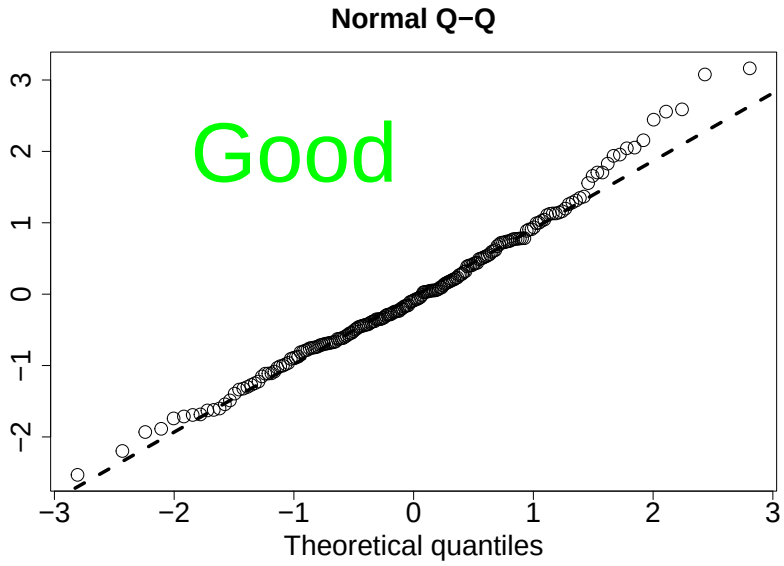
```
all.equal(cor(anscombe$x1,anscombe$y1),cor(anscombe$x2,anscombe$y2),
          cor(anscombe$x3,anscombe$y3),cor(anscombe$x4,anscombe$y4))
```

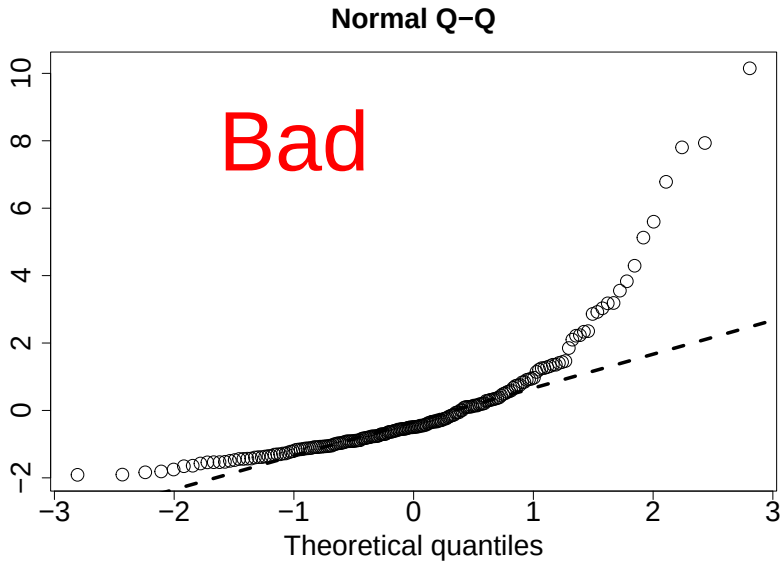
```
## [1] TRUE
```

Residual diagnostics: residuals vs. fitted



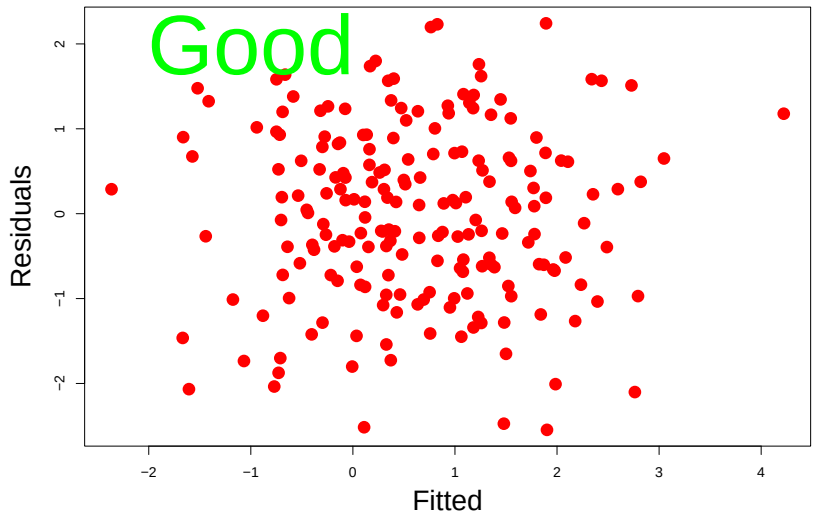
Residual diagnostics: normality with QQ-plot

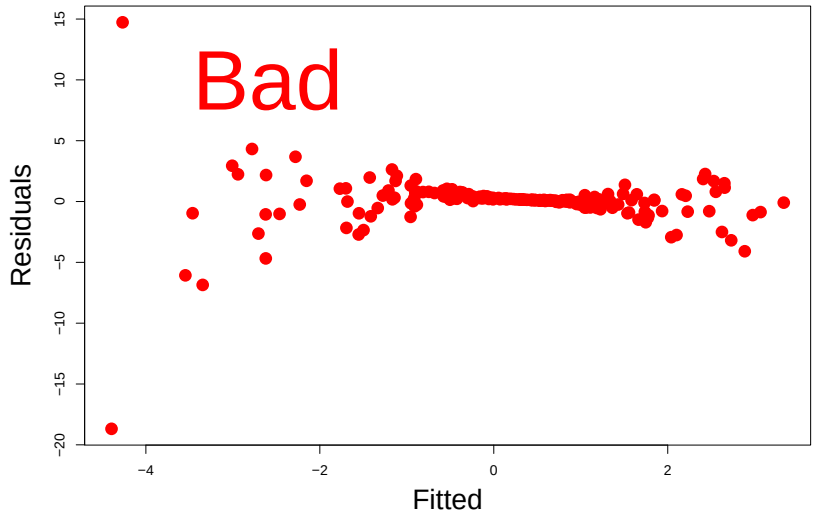




1. compare residuals to theoretical quantiles
2. points should follow the line

Residual diagnostics: constant variance





1. compare residuals to fitted
2. points should be randomly distributed

Data transformation

$$\mathrm{E}\{\mathrm{h}(y_i)|x_i\} = \alpha + \beta x_i \tag{5}$$

- $\mathrm{h}(\cdot)$ data transformation (e.g. $\sqrt{(\cdot)}$ or $\log(\cdot)$)
- To stabilize variance
-

To normalize residuals

Summary: occasions where violations arise

- outliers
- lack of linearity (partial residual plots)
- lack of constant variance (resid. vs. fitted)

More complex models: similar concepts apply

Next up

- bounded data types (e.g. counts)
- mean-variance
- lack of independence (due to e.g. study area, species, some kind of blocks)

Quick recap: Generalized Linear Model (GLM)

$$\mu_i = \mathrm{g}\{\mathrm{E}(y_i|x_i)\} = \alpha + \beta x_i \tag{6}$$

- $\mathrm{g}(\cdot)$ known link-function
- if $\mathrm{g}(\cdot) = \mathrm{h}(\cdot)$ approximately a GLM (Dunn and Smyth 2018 p. 232)
- MLE estimation
- No normality assumption

Assumptions

- Linearity
- Constant dispersion parameter ϕ for observations
- Independence
- Correct distribution & link function
- Correct variance function $V(\mu_i)$
- Mean-variance relationship (!)

Residuals

- Response residuals (i.e. $y_i - \hat{\mu}_i$) are insufficient
- We are after normally distributed residuals
  - for similar diagnostics

Generalized residuals

1) Working residuals
2) Pearson $r_P = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathrm{V}(\hat{\mu}_i)}}$
3) Anscombe (Anscombe 1953)

$$r_A = \frac{A(y_i) - A(\hat{\mu}_i)}{A'(\hat{\mu}_i)\sqrt{V(\hat{\mu}_i)}}, \qquad \text{where } A(\hat{\mu}_i) = \int_{-\infty}^{\hat{\mu}_i} \frac{1}{\mathrm{V}(t)^{1/3}} dt$$

4) Deviance $r_D = \mathrm{sgn}(y_i - \hat{\mu}_i)\sqrt{\mathrm{d}(y_i; \hat{\mu}_i)}$
5) Variance stabilizing (Pierce and Schafer 1986)
6) Randomized quantile residual

Residual distribution

- for normal, response residual = pearson = Anscombe = deviance
- approximately normally distributed (in large samples & large mean)
- often non-normal anyway (especially for small samples and discrete data)

Residuals for Poisson with log link-function

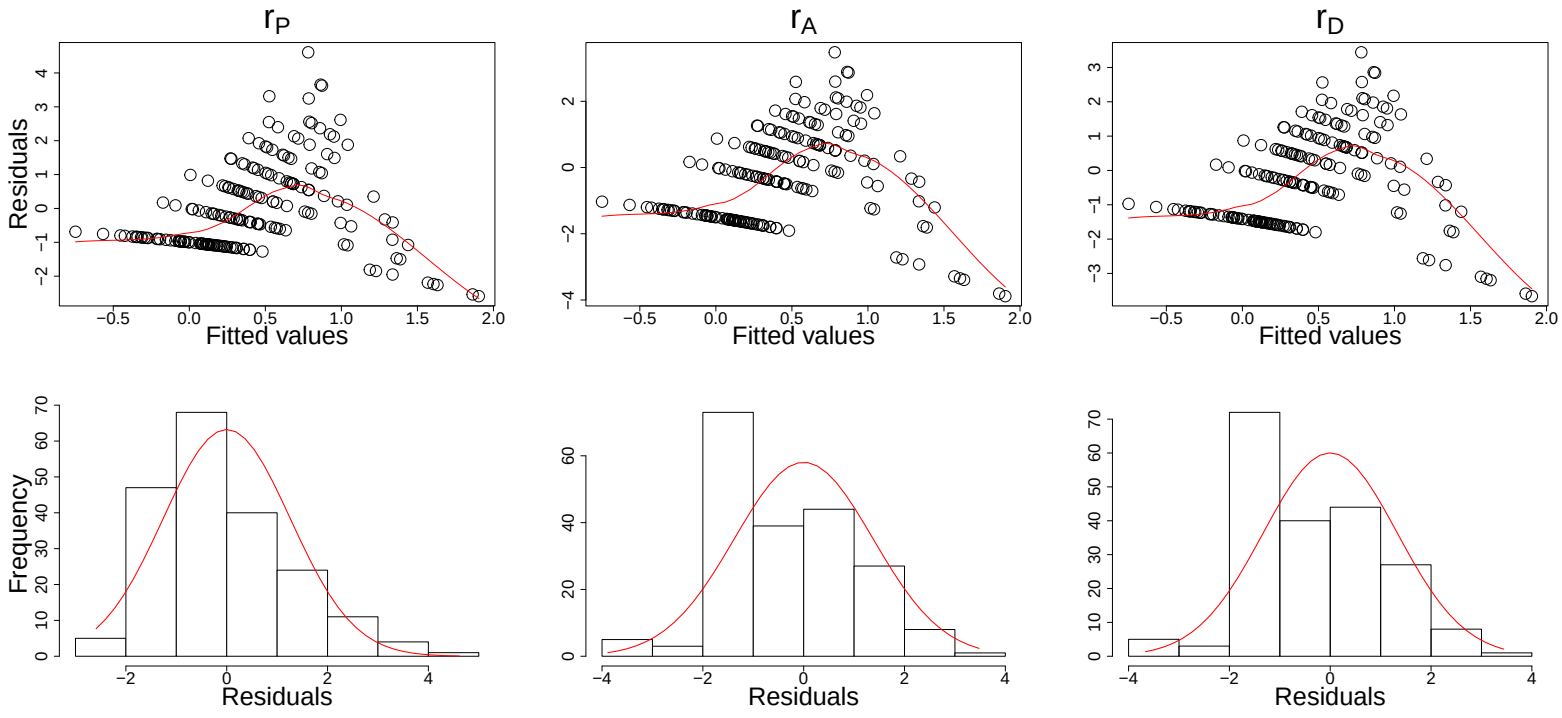
$$V(\alpha + \beta x_i) = g[\mathsf{E}(y_i|x_i)\}] = \mu_i = \exp(\alpha + \beta x_i)$$

2. $r_P = \frac{y_i - \exp(\alpha+\beta x_i)}{\sqrt{\exp(\alpha+\beta x_i)}}$
3. $r_A = \frac{3({y_i}^{2/3} - \exp(\alpha+\beta x_i)^{2/3})}{2\exp(\alpha+\beta x_i)^{1/6}}$
4. $r_D = \mathrm{sgn}\{y_i - \exp(\alpha+\beta x_i)\}\sqrt{2[y_i \log\{y_i/\exp(\alpha+\beta x_i)\} - y_i + \exp\{\alpha+\beta x_i\}]}$

Simulation: Poisson with log-link function

```
n <- 200
alpha <- 0.5
beta <- 0.5;beta2<--1
x <- rnorm(n = n)
mu <- exp(alpha + beta*x + beta2*x^2)
y <- rpois(n = n, lambda = mu)
mod <- glm(y~x, family = "poisson")
mu <- predict.glm(mod, type = "response")
```

Example: Poisson residuals



Why deviance residuals

- Converges faster to approximate normality

  (Dunn and Smyth 2018, Cox and Snell 1968)
- Otherwise, adjusted deviance residual
- Good for small samples (Pierce and Schafer 1986)
  - e.g. Poisson $\mu_i^{-0.5}/6$

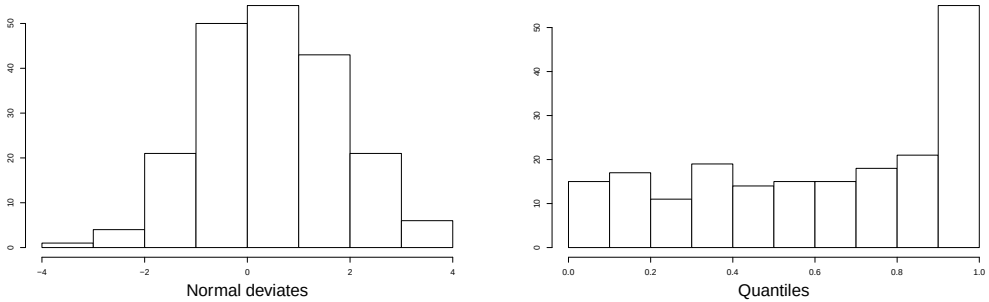
Still inappropriate for discrete data and small samples

Randomized Quantile residual (Dunn and Smyth 1996)

- Gold standard residual
- Better suited for small samples and discrete data types
- Exactly normally distributed

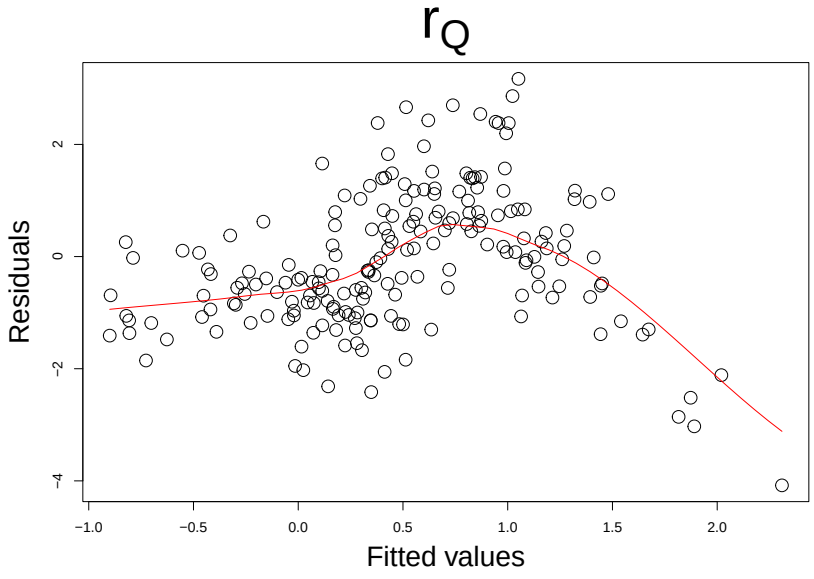
Continuous

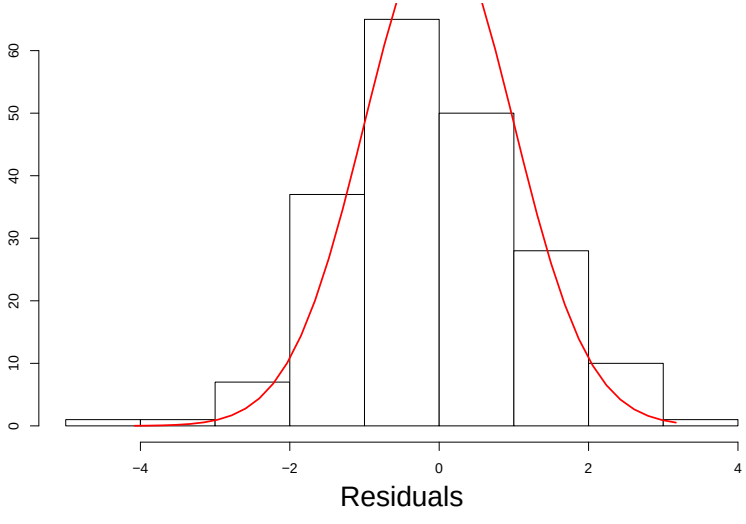
$$r_Q = \Phi^{-1}\left\{\mathcal{F}\left(y_i; \hat{\mu}_i, \hat{\phi}\right)\right\} \tag{7}$$



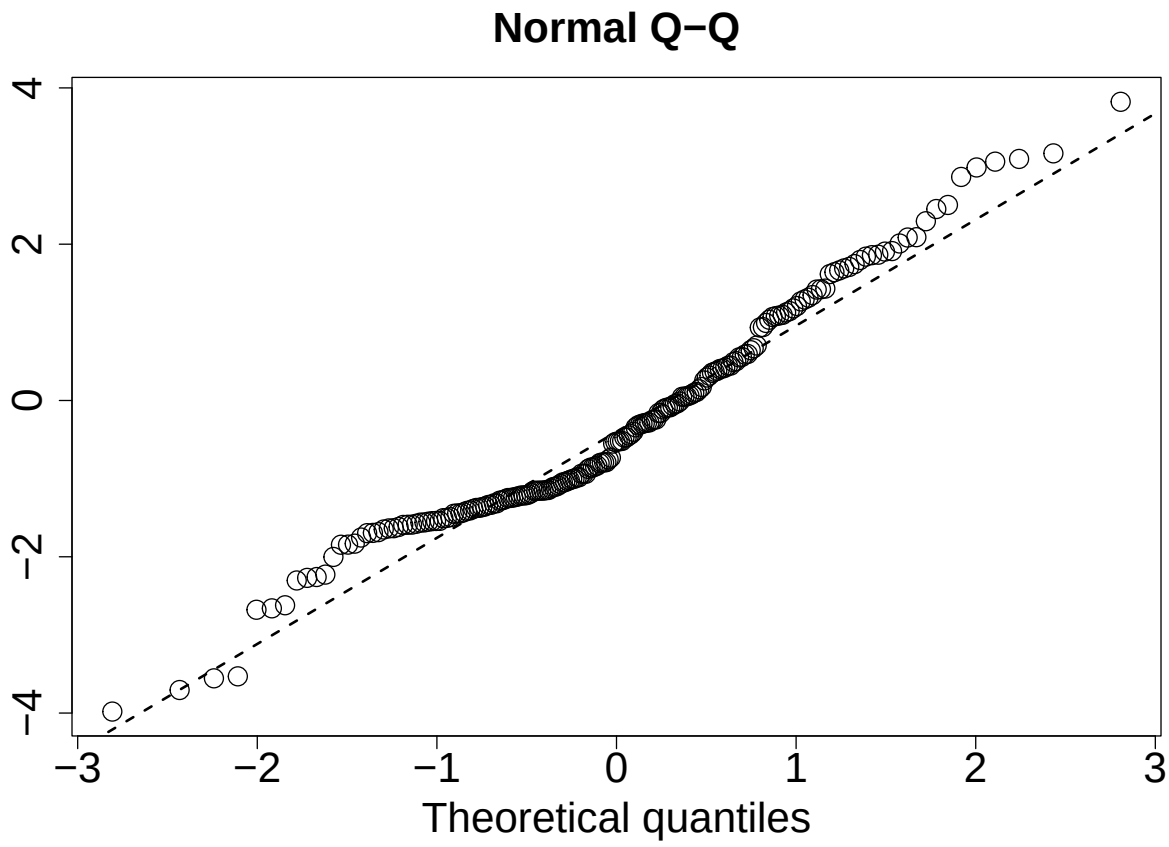
- Discrete data is mapped to a continuous space

Example: Poisson residuals (non-linearity)





Example: Poisson residuals QQ-plot (non-linearity)

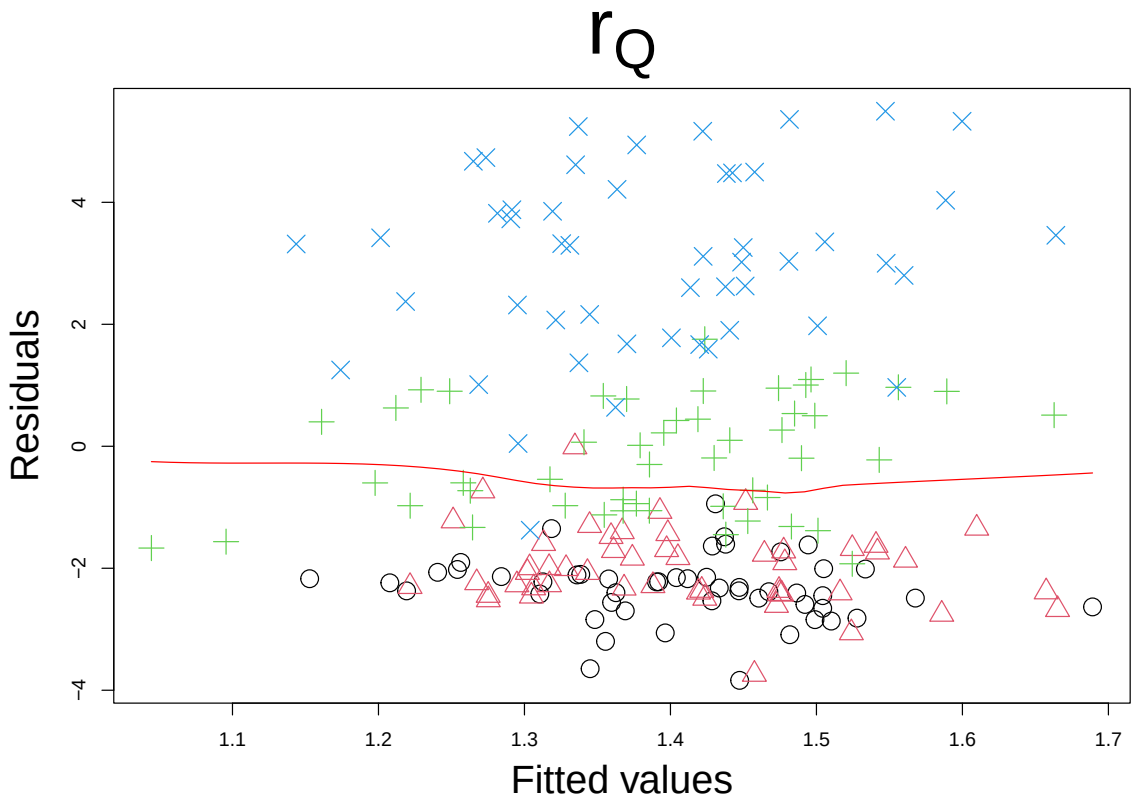


- Next up: pseudoreplication/grouping

Simulation: grouping of errors

```
n <- 200
ngroups <- 4
alpha <- 0.5
beta <- -1
x <- rnorm(n, sd = 0.2)

fac<-rep(1:ngroups,each=n/ngroups)
e <- seq(from=-2,to=2,length.out=ngroups)
mu <- exp(alpha + beta*x + e[fac])
y <- rpois(n = n, lambda = mu)
```

Example: Poisson residuals (grouping)



Generalized Linear Mixed-effects models

- Relax assumption: independence of errors

$$\mu_{ij} = \mathrm{g}\{\mathrm{E}(y_{ij}|x_i, b_j)\} = \alpha + \beta x_i + z_{ij} b_j, \qquad \text{where } b \sim \mathcal{N}(0, \Sigma) \tag{8}$$

- $j = 1 \ldots p$ groups
- e.g. study area, species, individual, blocks
- Σ accounts for grouping
- get prediction for $b_j = \mathrm{E}(b_j|y_i) = \hat{b}_j$

GLMM residuals

Conditional

$$\hat{\mu}_{ij} = \mathrm{g}\{\mathrm{E}(y_{ij}|x_i, \hat{b}_j)\} = \hat{\alpha} + \hat{\beta}x_i + z_{ij}\hat{b}_j \tag{9}$$

Unconditional

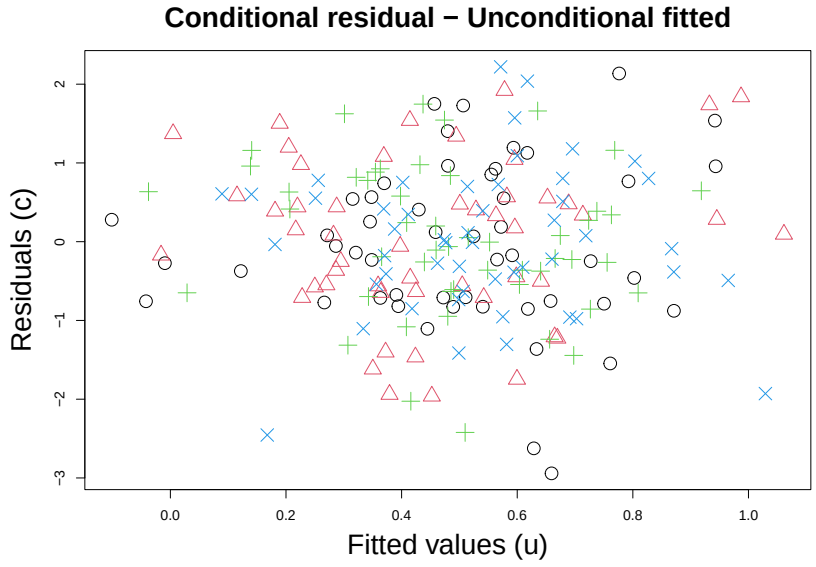
$$\hat{\mu}_{ij} = \mathrm{g}\{\mathrm{E}(y_{ij}|x_i)\} = \hat{\alpha} + \hat{\beta}x_i \tag{10}$$

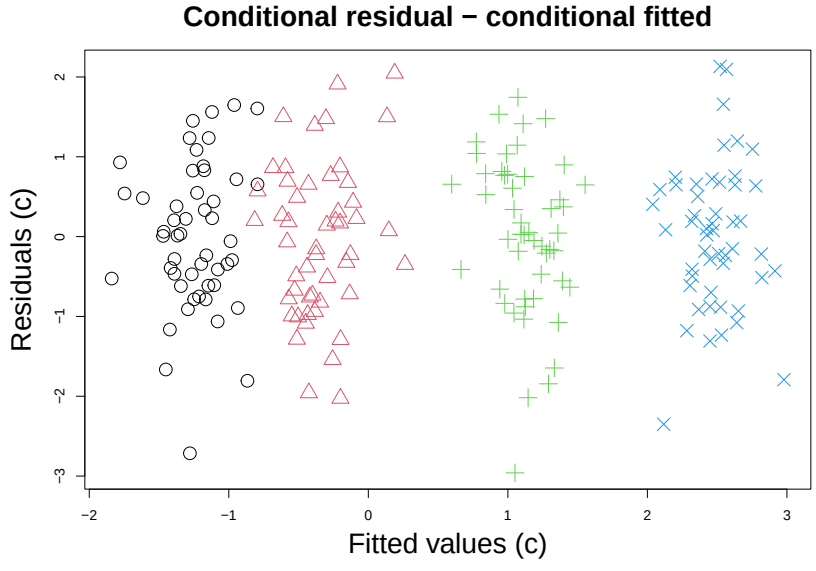
How do we calculate the residual?

- should we condition on the predicted random-effect?
- simulate from conditional distribution?

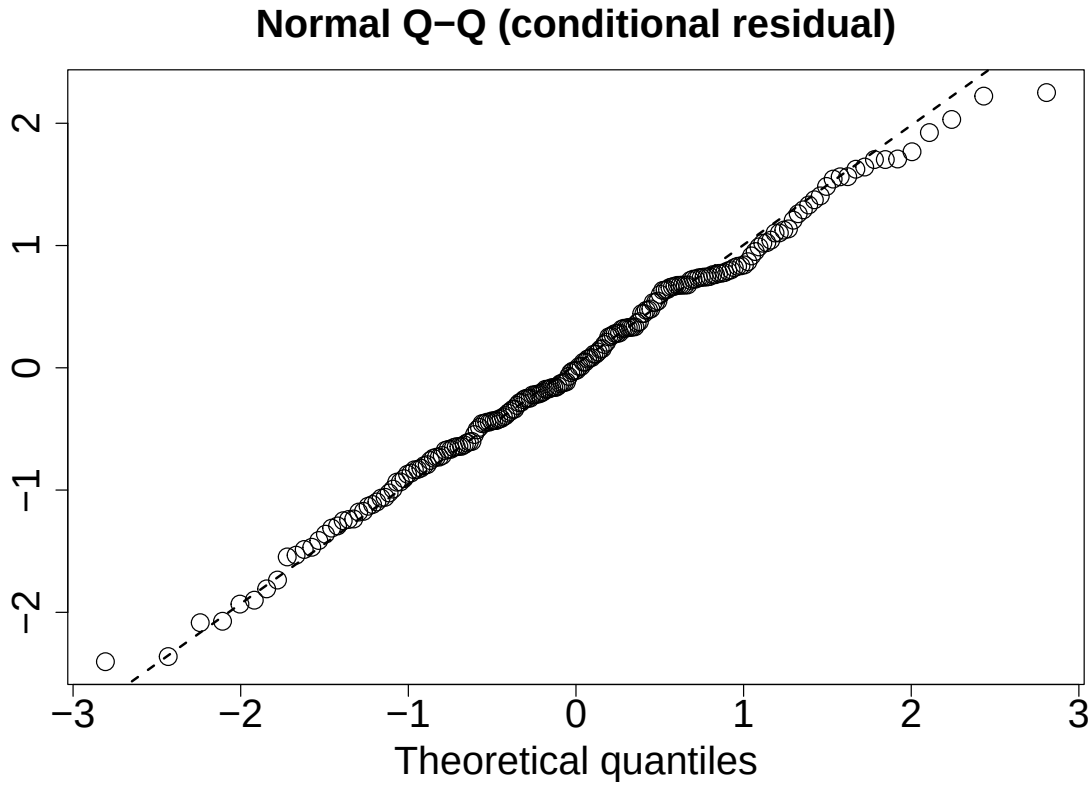
i.e. a range of options

Residual diagnostics: Poisson residuals (grouping)





Residual distributional assumptions: Poisson residuals



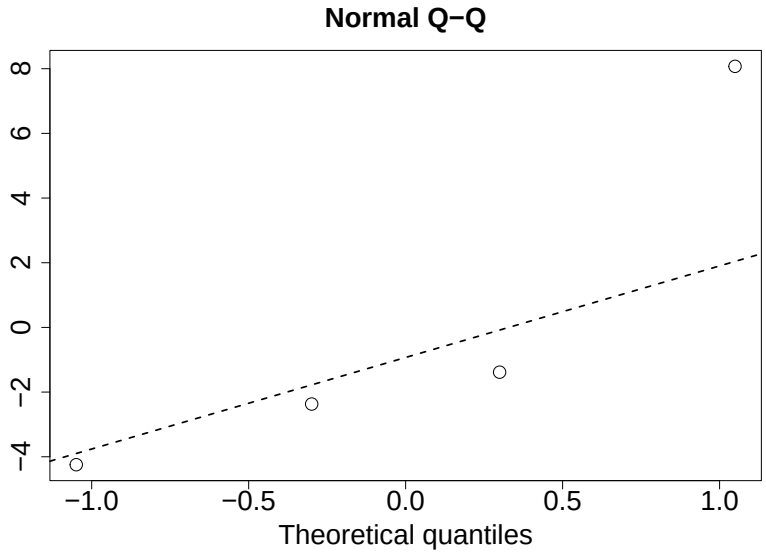
GLMM: checking random-effect assumptions

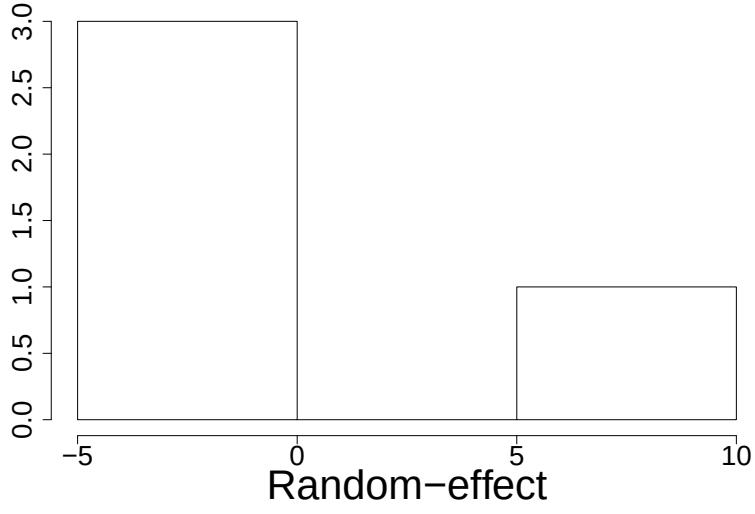
- random-effect is a type of residual
- $\widehat{b}_j$ is an estimate of the mean or mode of $p(b_j|y_i)$
- we treat $\widehat{b}_j$ as a sample of the random-effect distribution
- so we check assumptions (marginal normality, constant variance, independence, no outliers)!
- difficult with small number of groups
- needs to be done for every random-effect

Simulation: GLMM (outlier)

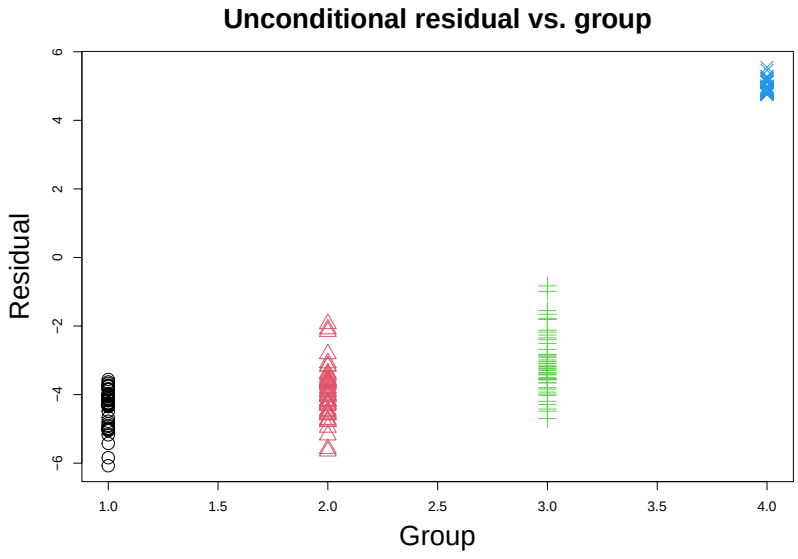
```
n <- 200
ngroups <- 4
alpha <- 0.5
beta <- -1
x <- rnorm(n, sd = 0.2)

fac<-rep(1:ngroups,each=n/ngroups)
e <- seq(from=-2,to=2,length.out=ngroups)
e[4] <- 10
mu <- exp(alpha + beta*x + e[fac])
y <- rpois(n = n, lambda = mu)
```

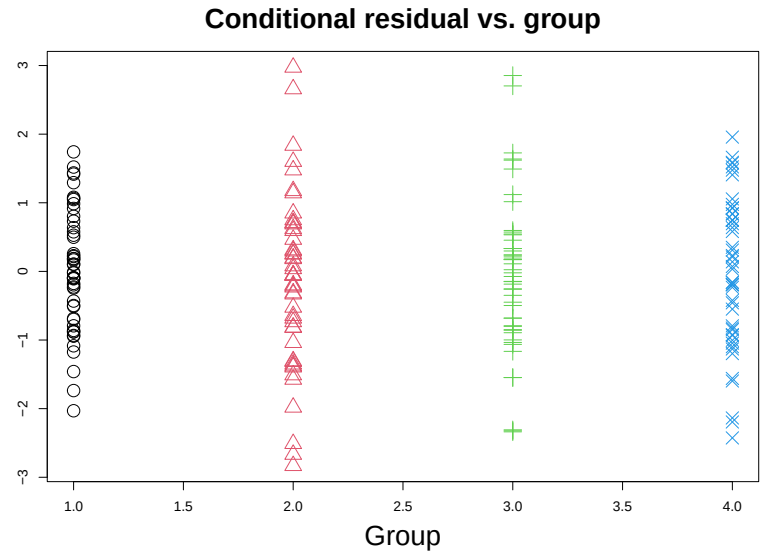
GLMM diagnostics





GLMM diagnostics

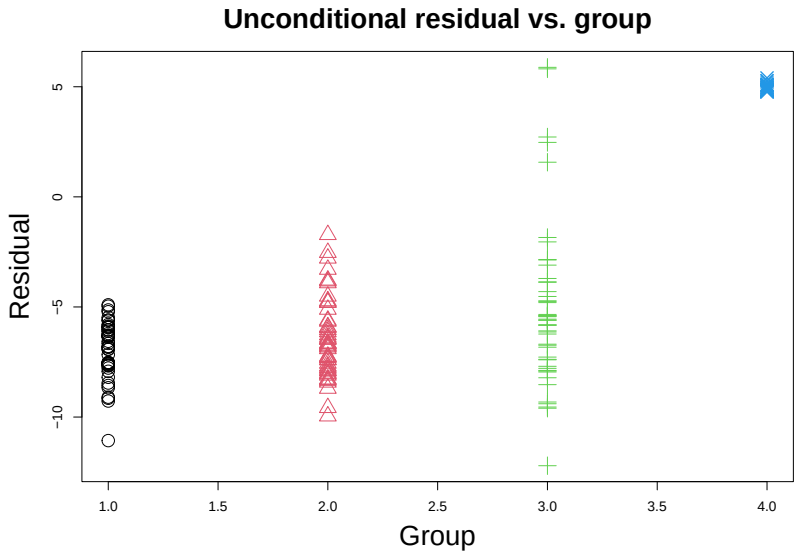


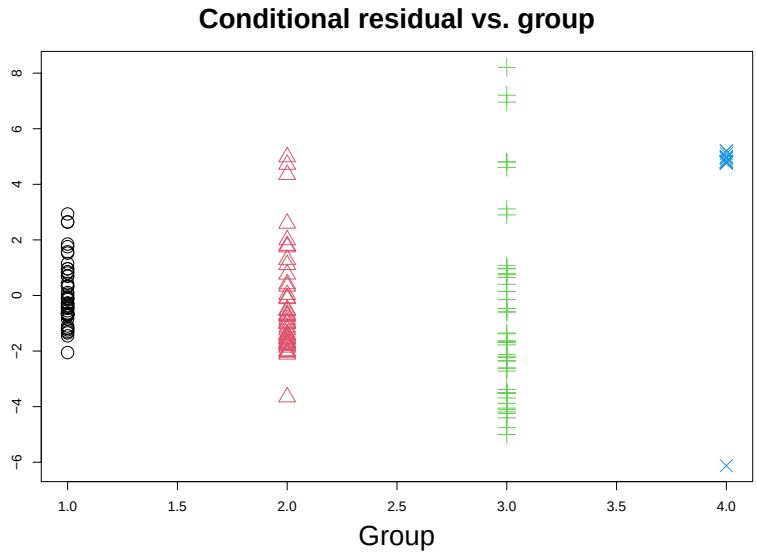


What if constant variance is violated

```
n <- 200
ngroups <- 4
alpha <- 0.5
beta <- -1
x <- rnorm(n, sd = 0.2)

fac<-rep(1:ngroups,each=n/ngroups)
e <- seq(from=-2,to=2,length.out=ngroups)
e[4] <- 10
e2 <- MASS::mvrnorm(1,rep(0,n), diag(rep(c(1,2,3,4),
                                         each=n/ngroups)))
mu <- exp(alpha + beta*x + e[fac] + e2)
y <- rpois(n = n, lambda = mu)
```

GLMM diagnostics





Next up “Beyond”

- Some short notes on multivariate models

Multivariate analysis

$$Y = XB + \mathrm{E} \tag{11}$$

- same assumptions linear regression
- independence of rows i
- $\epsilon_i \sim \mathcal{N}(0, \Sigma)$
- i.e. Σ is a $p \times p$ covariance matrix
- Cholesky-transformed residual

Joint Species Abundance/Distribution Model

- columns are species (or individuals or..)

$$\mathrm{g}\{\mathrm{E}(y_{ij}|x_i, \epsilon_{ij})\} = \alpha_j + \beta_j x_i + \epsilon_{ij} \tag{12}$$

- **conditional** independence of species j
- residual checks same as for GLMMs
  - but many clusters
  - and many random-effects

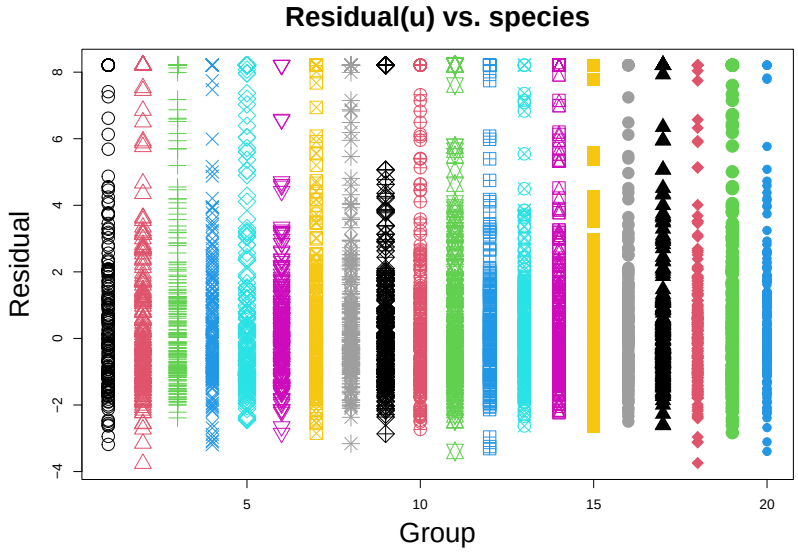
Generalized Linear Latent Variable Models

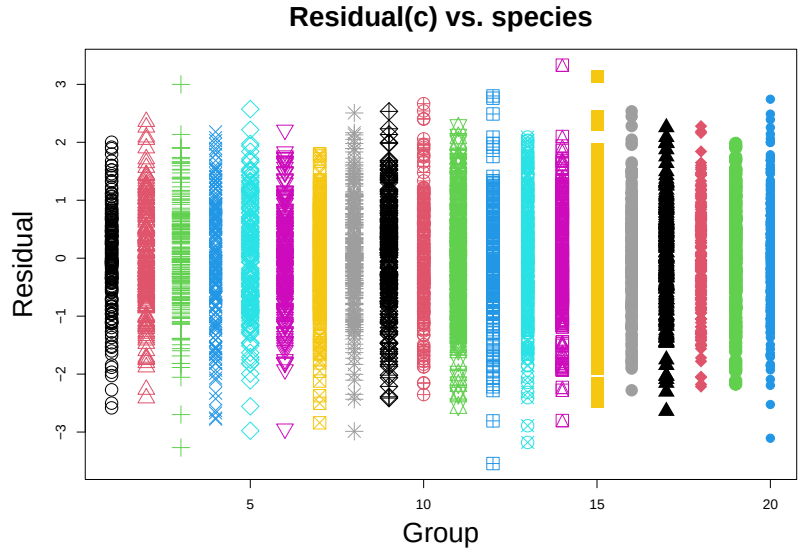
- Same model
- but instead $\text{rank}(\Sigma) = d \ll p$
- "factor analytic" or "latent variable" approach
- $\epsilon_{ij} = z_i^\top \gamma_j$ with $\Sigma = \gamma_j \gamma_j^\top$
- $z_i \sim \mathcal{N}(0, I_d)$

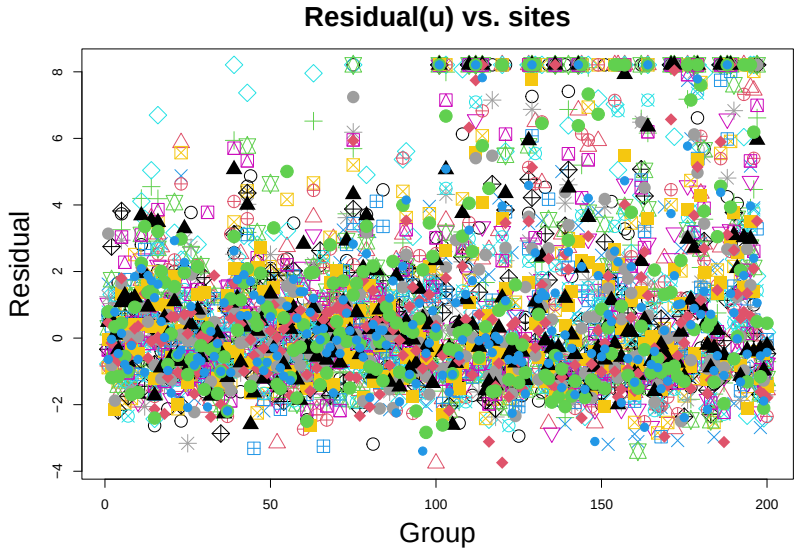
Simulation: GLLVM

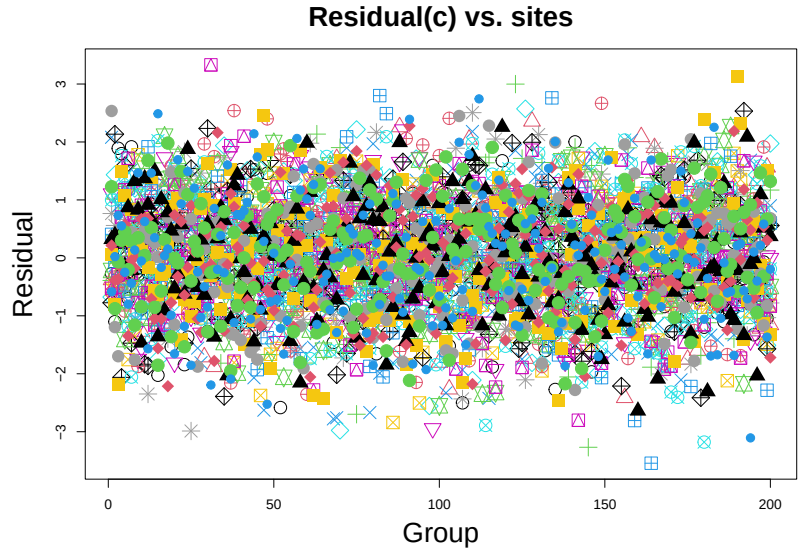
```
n <- 200
nspecies <- 20
d <- 2
set.seed(1);alpha <- runif(nspecies)
alpha <- matrix(alpha,ncol=nspecies,nrow=n,byrow=T)

set.seed(1);z <- MASS::mvrnorm(n/2,rep(0,d),diag(d))
set.seed(1);z2 <- MASS::mvrnorm(n/2,rep(0,d),diag(10,d))
z <- rbind(z,z2)
set.seed(1);gamma <- matrix(runif(nspecies*d),ncol=d)
gamma[upper.tri(gamma)] <- 0
e <- z%*%t(gamma)
mu <- exp(alpha+e)
set.seed(1);y <- rpois(n = n*nspecies, lambda = mu)
y <- matrix(y,ncol=nspecies,nrow=n)
```

GLLVM diagnostics



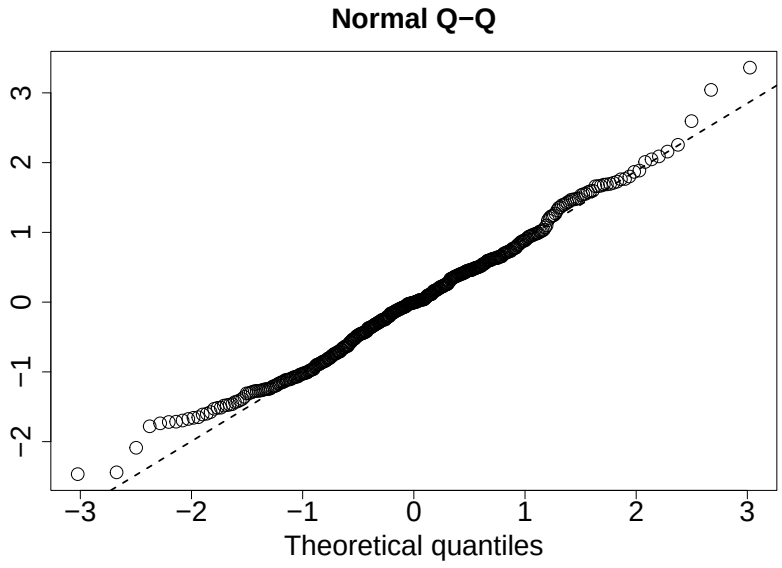


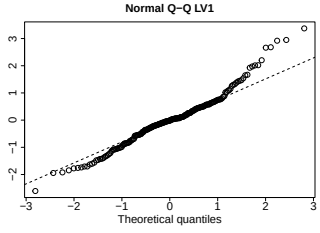


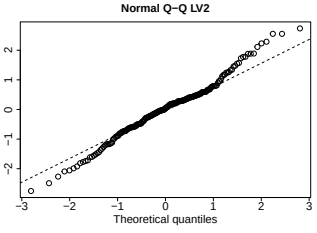


GLLVM diagnostics

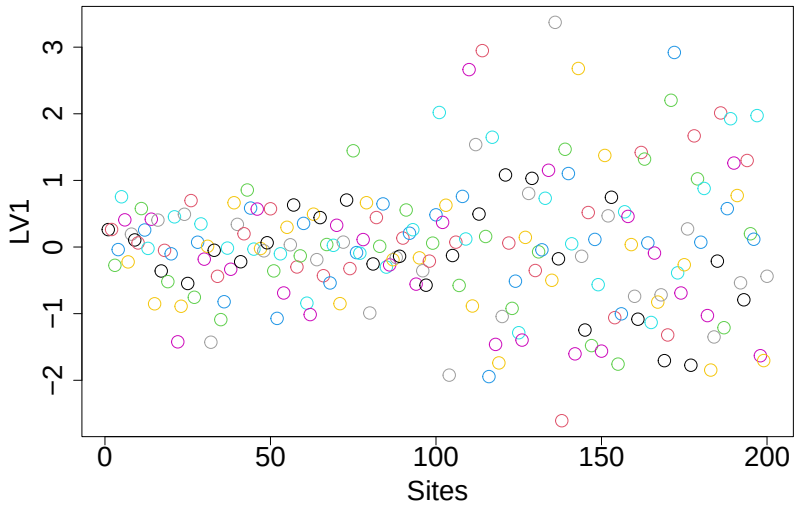
- Now we have many random-effects
- Which should come from the same distribution

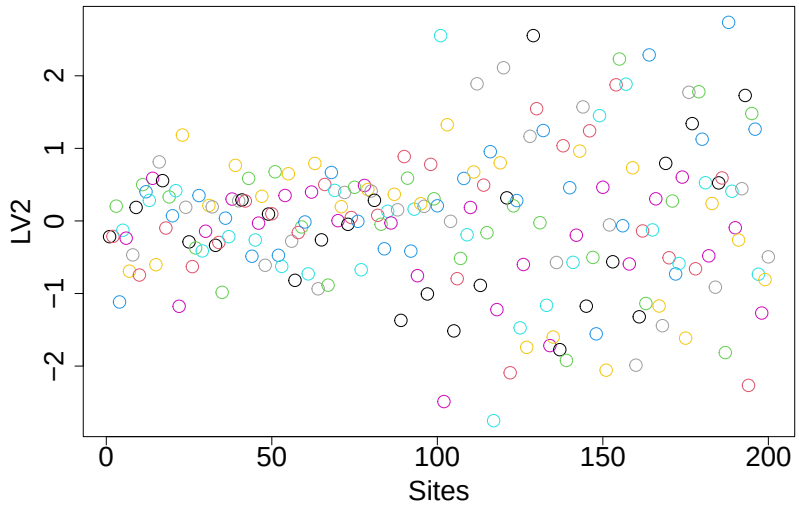






LV versus sites





Finally

- here I used quantile residuals for glms, glmms, and gllvms diagnostics
- also suitable for more model types (e.g. GAMs, occupancy)
- more complex model = more assumptions to check