Untitled

J Lacasa

2024-10-17

Libraries

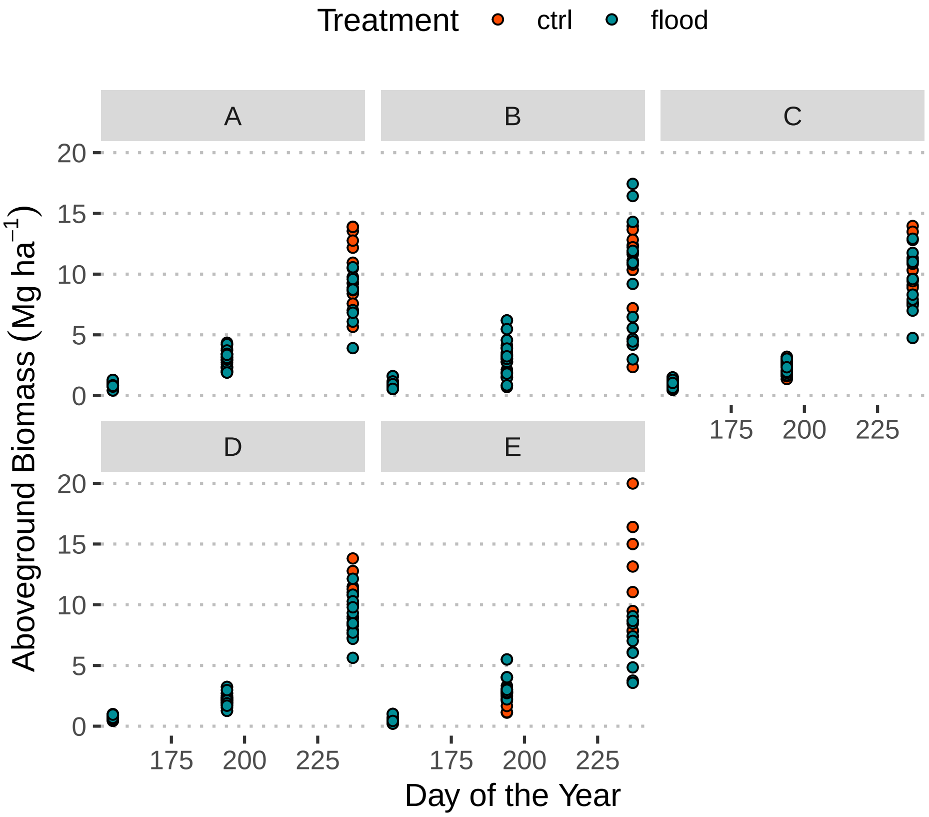
```
library(tidyverse) # data wrangling, data viz
library(mgcv) # fitting generalized additive models
library(latex2exp) # math notation in plots
library(ggpubr) # plot aesthetics
```

Loading the data

```
# dd_finalproj <- read.csv("../classes/data/dd_finalproj.csv")
url <- "https://raw.githubusercontent.com/jlacasa/stat705_fall2024/main/classes/data/dd_finalproj.csv"
dd_finalproj <- read.csv(url)
dd_finalproj$doy_f <- factor(dd_finalproj$doy)
```

Exploratory Data Analysis

```
dd_finalproj %>%
  ggplot(aes(doy, fitted))+
  geom_point(aes(y = agb_g, fill = trt), shape = 21)+
  facet_wrap(~species)+
  scale_fill_manual(values = c("#FC4C02", "#008E97"))+
  labs(y = expression(Aboveground~Biomass~(Mg~ha^{-1})),
       x= "Day of the Year",
       color = "Treatment",
       fill = "Treatment")+
  theme_pubclean()+
  theme(aspect.ratio = 1)
```



Model Fitting

Firstly, fit a simple model

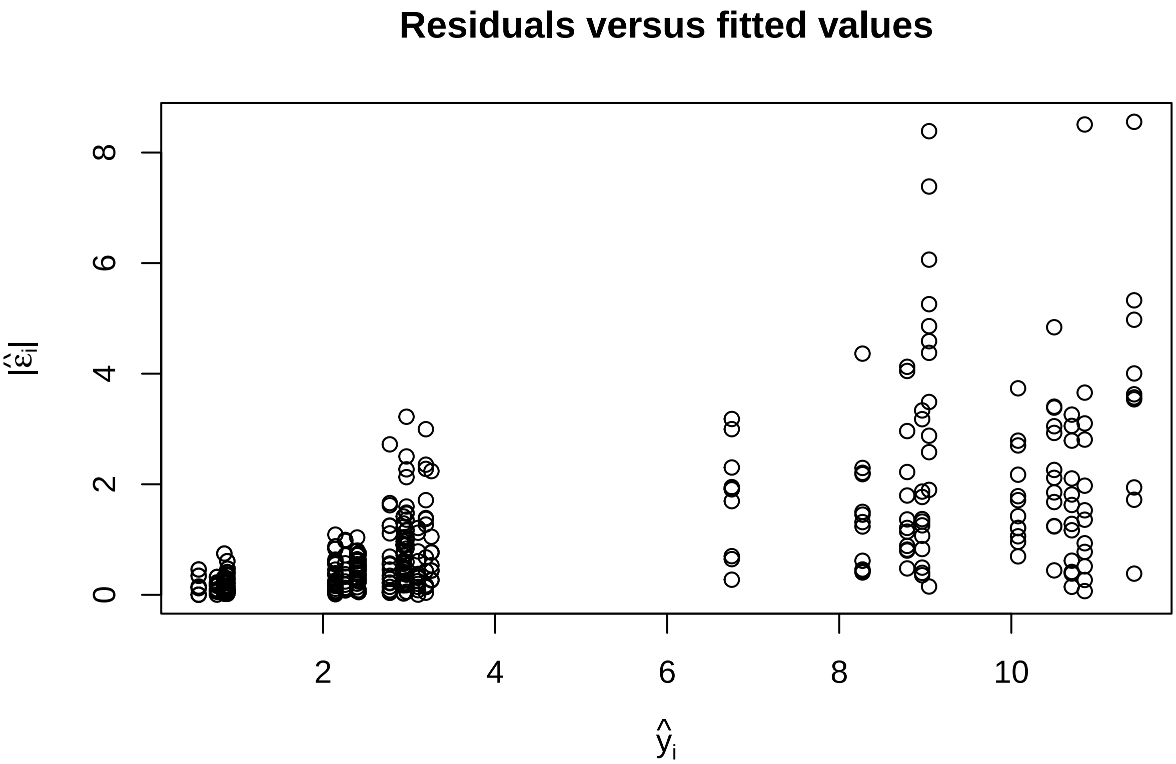
$$y_{ijkl} \sim N(\mu_{ijk}, \sigma^2)$$

$$\mu_{ijk} = \beta_0 + \tau_i + \rho_j + \alpha_k + (\tau\rho)_{ij} + (\tau\alpha)_{ik} + (\rho\alpha)_{ik} + (\tau\rho\alpha)_{ijk},$$

where y_{ijkl} is the observation of aboveground biomass (in g) for the ith treatment, jth species, kth moment (i.e., time), and lth repetition, that arises from a Normal distribution with mean mu_{ijk} and variance $sigma^2$. The β_0 is the general intercept, τ_i is the effect of the ith treatment, ρ_j is the effect of the jth species, and $(\tau\rho)_{ij}$, $(\tau\alpha)_{ik}$, and $(\rho\alpha)_{ik}$ are the two-way interactions between the factors mentioned above, and $(\tau\rho\alpha)_{ijk}$ is the three-way interaction of the factors mentioned above. Note that there is a single variance for all observations.

Model diagnostics

```
plot(m1$fitted.values, abs(residuals(m1, type = "deviance")),
     xlab = TeX("$\\hat{y}_i$"),
     ylab = TeX("|$\\hat{\\epsilon}_i$|"),
     main = "Residuals versus fitted values",
     sub = "Note that the errors increase together with the means")
```



Note that the errors increase together with the means

Constant variance

```
lm(log(agb_g) ~ species*trt*doy_f, data = dd_finalproj)
```

```
##
## Call:
## lm(formula = log(agb_g) ~ species * trt * doy_f, data = dd_finalproj)
##
## Coefficients:
##                (Intercept)                    speciesB
##                 -1.514e-01                  -6.419e-02
##                   speciesC                    speciesD
##                 -3.885e-02                  -1.388e-01
##                   speciesE                    trtflood
##                 -5.354e-01                   8.716e-15
##                   doy_f194                    doy_f237
##                  1.194e+00                   2.468e+00
##          speciesB:trtflood           speciesC:trtflood
##                 -9.031e-15                  -9.126e-15
##          speciesD:trtflood           speciesE:trtflood
##                 -5.422e-15                  -9.659e-15
##          speciesB:doy_f194           speciesC:doy_f194
##                 -1.057e-02                  -1.565e-01
##          speciesD:doy_f194           speciesE:doy_f194
##                 -1.646e-01                   4.488e-01
##          speciesB:doy_f237           speciesC:doy_f237
##                  5.510e-02                   7.766e-02
##          speciesD:doy_f237          speciesE:doy_f237
##                  1.135e-01                  5.849e-01
##          trtflood:doy_f194          trtflood:doy_f237
##                  6.994e-02                 -2.364e-01
## speciesB:trtflood:doy_f194 speciesC:trtflood:doy_f194
##                 -2.277e-03                 -5.797e-02
## speciesD:trtflood:doy_f194 speciesE:trtflood:doy_f194
##                 -2.403e-02                  1.197e-01
## speciesB:trtflood:doy_f237 speciesC:trtflood:doy_f237
##                 -2.553e-02                  2.229e-02
## speciesD:trtflood:doy_f237 speciesE:trtflood:doy_f237
##                  1.190e-01                 -2.714e-01
```

Model fitting II

The first, simpler, model (i.e., `m1`) does not seem to have constant variance. We can follow two routes: (1) transform the response and keep the assumptions, or (2) keep the data as is and change our model assumptions.

Following the second option, we can model

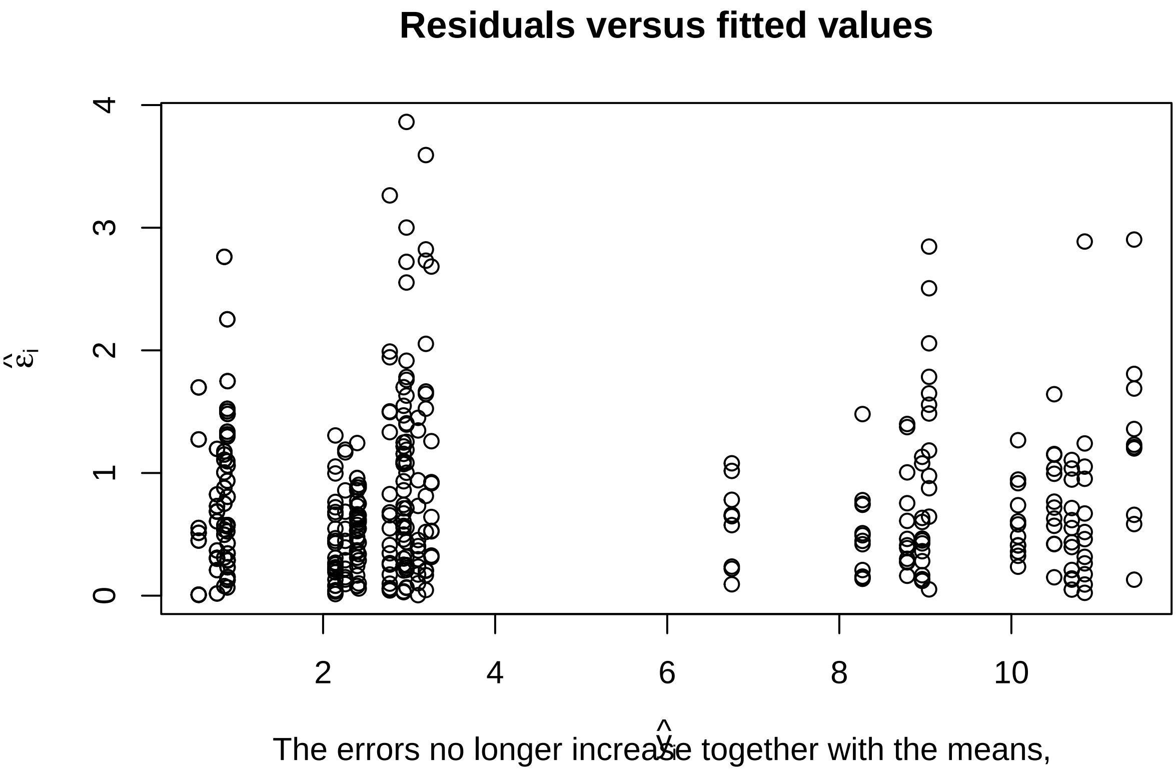
$$y_{ijkl} \sim N(\mu_{ijk}, \sigma^2_k)$$
$$\mu_{ijk} = \beta_0 + \tau_i + \rho_j + \alpha_k + (\tau\rho)_{ij} + (\tau\alpha)_{ik} + (\rho\alpha)_{ik} + (\tau\rho\alpha)_{ijk},$$

where y_{ijkl} is the observation of aboveground biomass (in g) for the ith treatment, jth species, kth moment (i.e., time), and lth repetition, that arises from a Normal distribution with mean mu_{ijk} and variance at time k, $sigma^2_k$. The β_0 is the general intercept, τ_i is the effect of the ith treatment, ρ_j is the effect of the jth species, and $(\tau\rho)_{ij}$, $(\tau\alpha)_{ik}$, and $(\rho\alpha)_{ik}$ are the two-way interactions between the factors mentioned above, and $(\tau\rho\alpha)_{ijk}$ is the three-way interaction of the factors mentioned above. Note that the variance is a function of time.

```
m2 <- gam(list(agb_g ~ species*trt*factor(doy),
               ~ doy),
          family = gaulss(),
          data = dd_finalproj)
m2_fitted <- predict(m2)[,1]
m2_residuals <- residuals(m2, type = "deviance")
```

Model diagnostics II

```
plot(m2_fitted, abs(m2_residuals),
     xlab = TeX("$\\hat{y}_i$"),
     ylab = TeX("$\\hat{\\epsilon}_i$"),
     main = "Residuals versus fitted values",
     sub = "The errors no longer increase together with the means,
     but there seems to be different dispersions for different species")
```



The errors no longer increase together with the means,
but there seems to be different dispersions for different species

Constant variance

Model fitting III

The second, model (i.e., `m2`) fixes the issue of constant variance. However, the variance seems to be different depending on the groups.

We can thus further model σ^2:

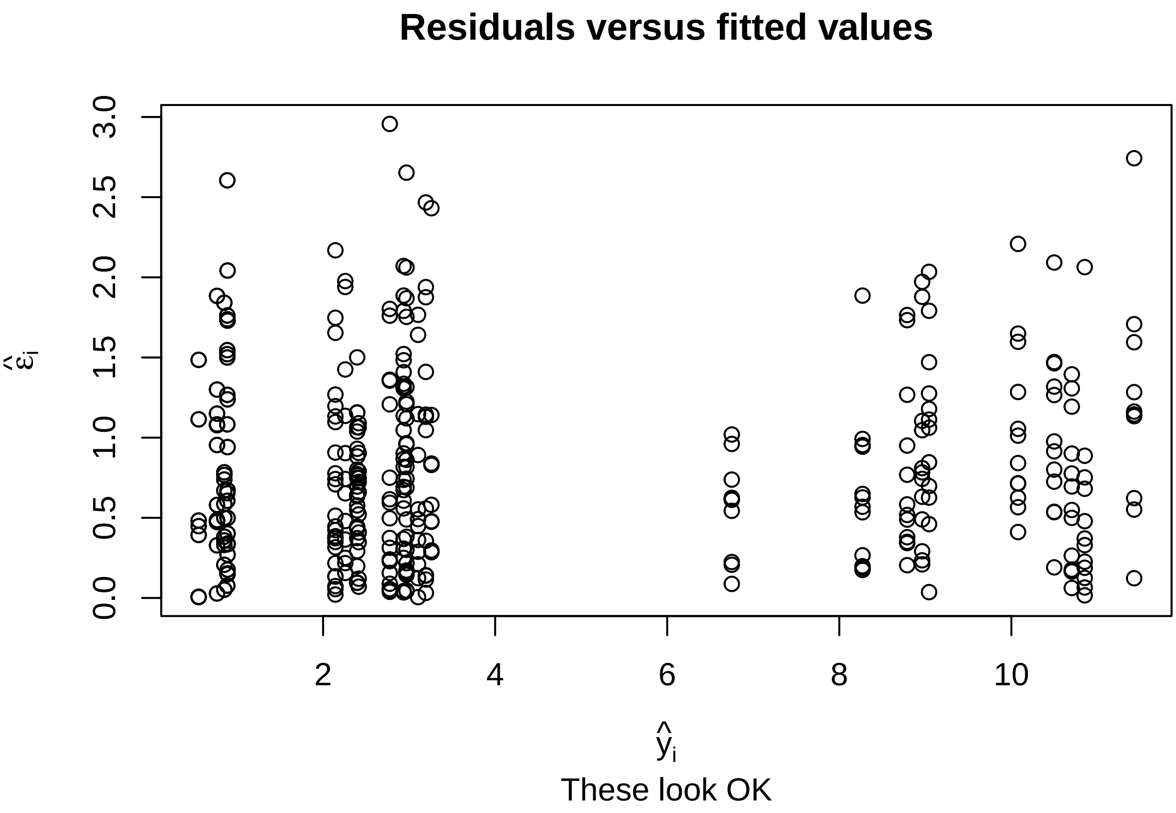
$$y_{ijkl} \sim N(\mu_{ijk}, \sigma^2_{jk})$$
$$\mu_{ijk} = \beta_0 + \tau_i + \rho_j + \alpha_k + (\tau\rho)_{ij} + (\tau\alpha)_{ik} + (\rho\alpha)_{ik} + (\tau\rho\alpha)_{ijk},$$

where y_{ijkl} is the observation of aboveground biomass (in g) for the ith treatment, jth species, kth moment (i.e., time), and lth repetition, that arises from a Normal distribution with mean mu_{ijk} and variance at time k for species j, $sigma^2_{jk}$. The β_0 is the general intercept, τ_i is the effect of the ith treatment, ρ_j is the effect of the jth species, and $(\tau\rho)_{ij}$, $(\tau\alpha)_{ik}$, and $(\rho\alpha)_{ik}$ are the two-way interactions between the factors mentioned above, and $(\tau\rho\alpha)_{ijk}$ is the three-way interaction of the factors mentioned above. Note that the variance is a function of time.

```
m3 <- gam(list(agb_g ~ species*trt*factor(doy),
               ~ doy + species),
          family = gaulss(),
          data = dd_finalproj)
m3_fitted <- predict(m3)[,1]
m3_residuals <- residuals(m3, type = "deviance")
```

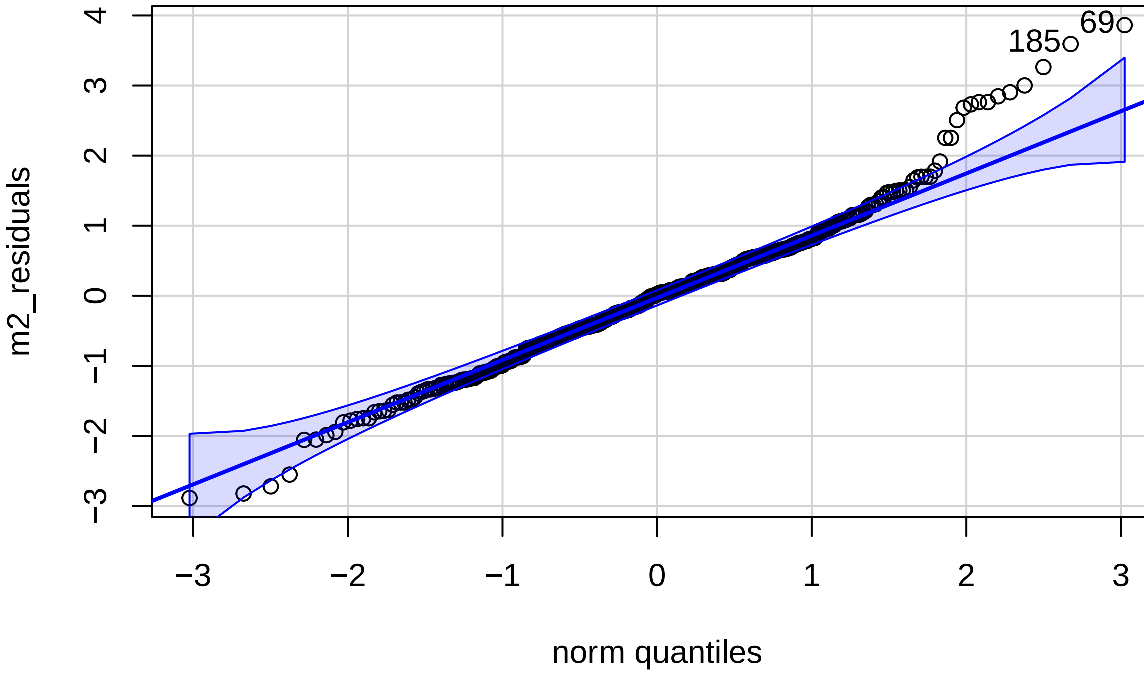
Model diagnostics III

```
plot(m3_fitted, abs(m3_residuals),
     xlab = TeX("$\\hat{y}_i$"),
     ylab = TeX("$\\hat{\\epsilon}_i$"),
     main = "Residuals versus fitted values",
     sub = "These look OK")
```



Constant variance

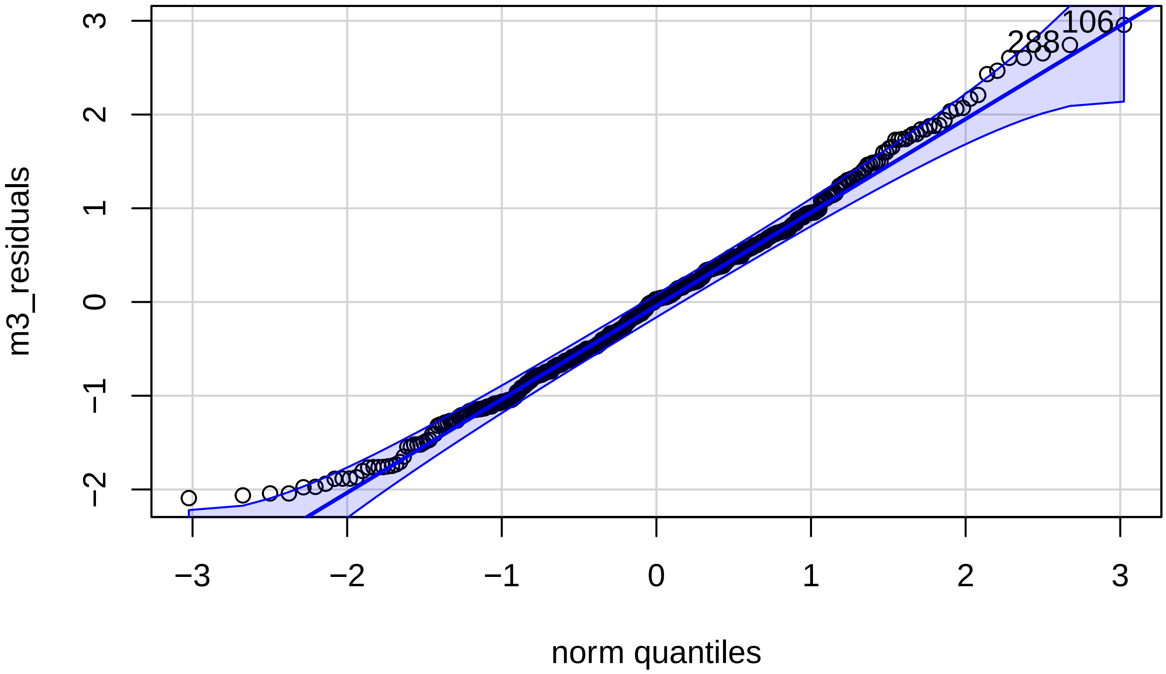
```
car::qqPlot(m2_residuals)
```



Normal distribution

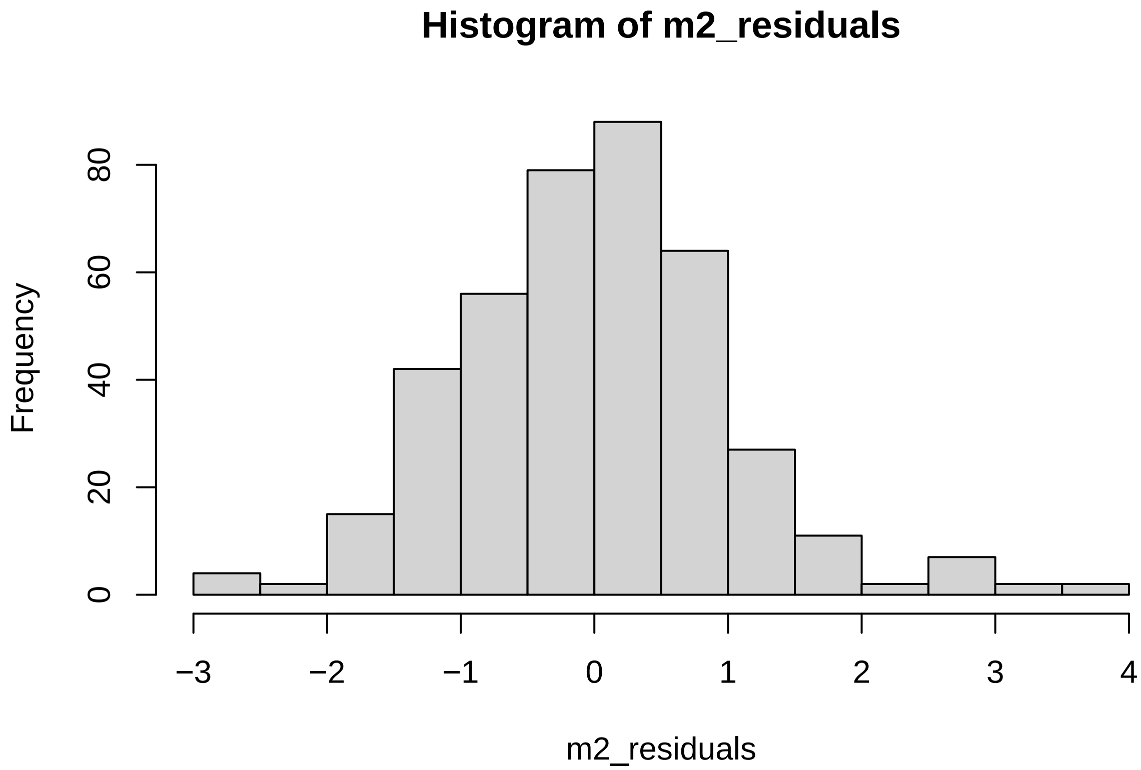
```
## [1]  69 185
```

```
car::qqPlot(m3_residuals)
```

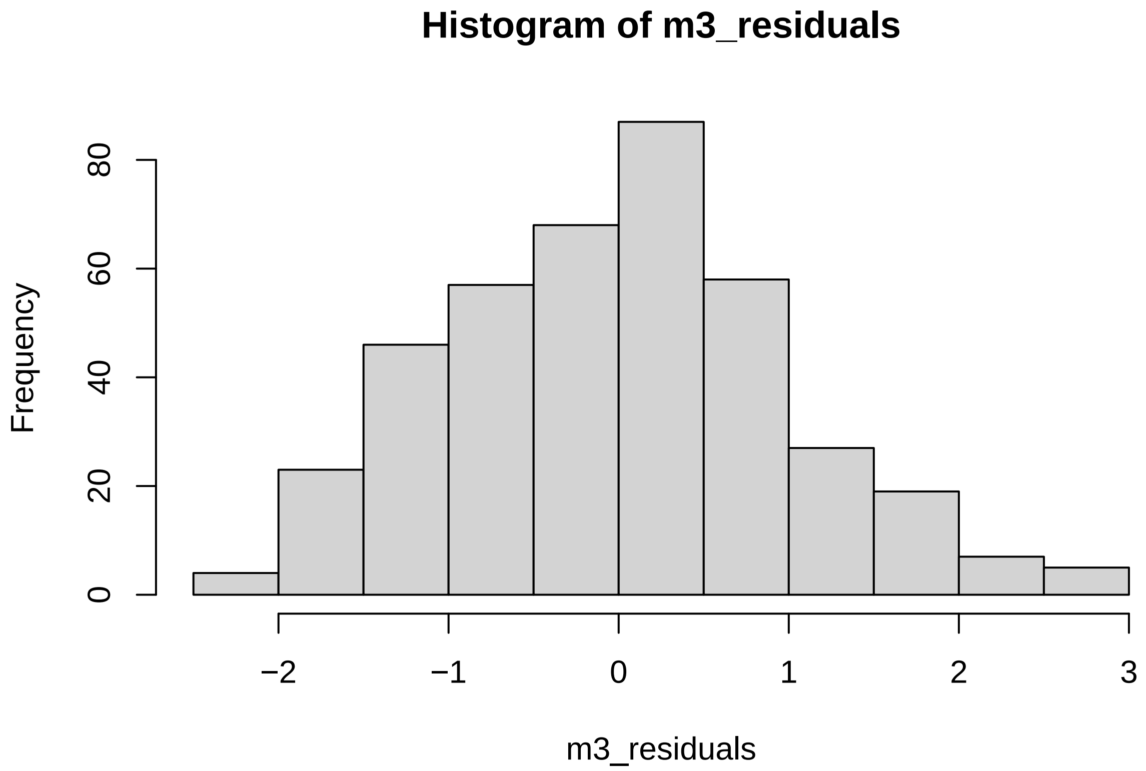


```
## [1] 106 288
```

```
hist(m2_residuals)
```



```
hist(m3_residuals)
```



AIC, BIC

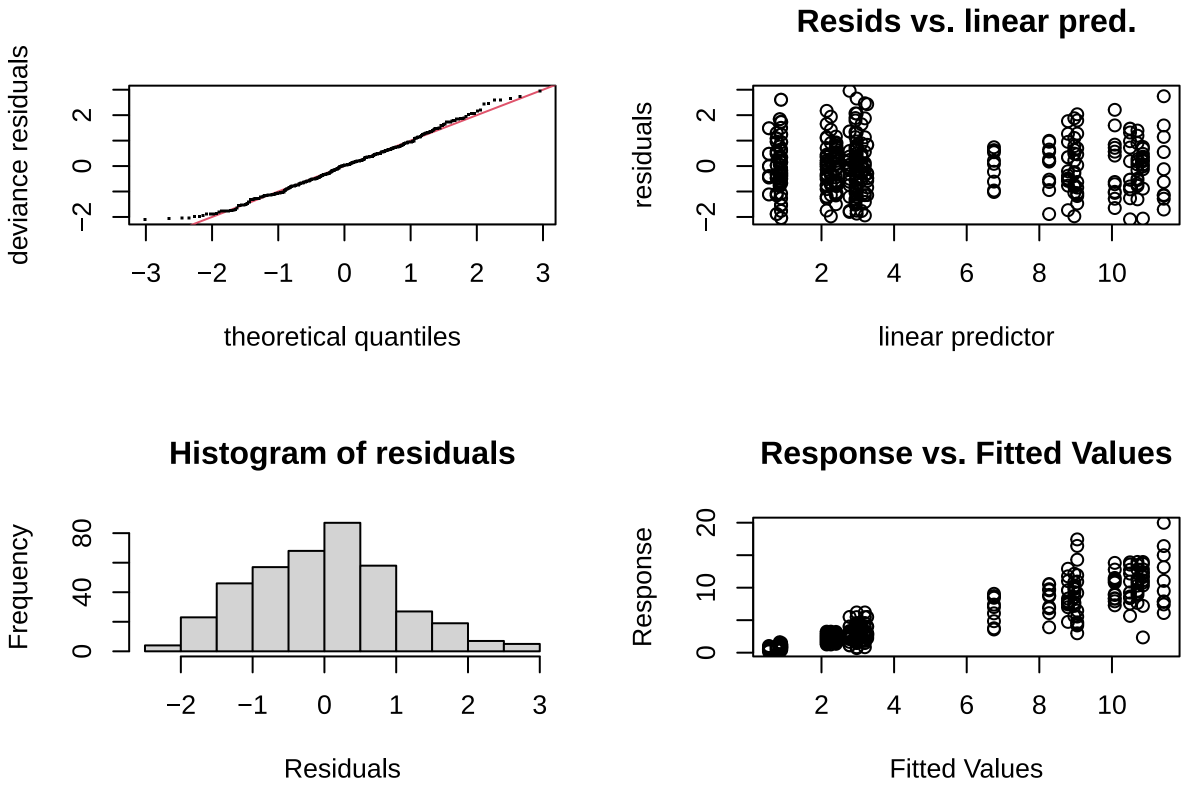
```
left_join(AIC(m1,m2,m3) %>% rowmodel_to_column("model"),
          BIC(m1,m2,m3) %>% rownames_to_column("model"))
```

```
## Joining with ‘by = join_by(model, df)‘
```

```
##   model df      AIC      BIC
## 1    m1 31 1594.928 1718.741
## 2    m2 32 1101.115 1228.922
## 3    m3 36 1034.997 1178.780
```

Other checks

```
gam.check(m3)
```



```
##
## Method: REML   Optimizer: outer newton
## Model required no smoothing parameter selectionModel rank =  36 / 36
```

Inference

```
summary(m3)
```

Summary

```
##
## Family: gaulss
## Link function: identity logb
##
## Formula:
## agb_g ~ species * trt * factor(doy)
## ~doy + species
##
## Parametric coefficients:
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        8.900e-01  6.681e-02  13.321  < 2e-16 ***
## speciesB                          -3.767e-02  1.347e-01  -0.280 0.779799
## speciesC                         -3.053e-03  9.495e-02  -0.032 0.974351
## speciesD                         -1.224e-01  8.317e-02  -1.471 0.141218
## speciesE                         -3.361e-01  1.345e-01  -2.498 0.012475 *
## trtflood                          1.820e-17  9.448e-02   0.000 1.000000
## factor(doy)194                    2.047e+00  1.549e-01  13.215  < 2e-16 ***
## factor(doy)237                    9.608e+00  6.711e-01  14.316  < 2e-16 ***
## speciesB:trtflood                 4.215e-17  1.906e-01   0.000 1.000000
## speciesC:trtflood                -3.364e-18  1.343e-01   0.000 1.000000
## speciesD:trtflood                -2.778e-17  1.176e-01   0.000 1.000000
## speciesE:trtflood                -4.214e-17  1.902e-01   0.000 1.000000
## speciesB:factor(doy)194           7.061e-02  3.148e-01   0.224 0.822546
## speciesC:factor(doy)194          -5.368e-01  2.201e-01  -2.439 0.014741 *
## speciesD:factor(doy)194          -6.710e-01  1.922e-01  -3.490 0.000483 ***
## speciesE:factor(doy)194           1.749e-01  2.828e-01   0.618 0.536332
## speciesB:factor(doy)237           3.923e-01  1.371e+00   0.286 0.774820
## speciesC:factor(doy)237           2.076e-01  9.540e-01   0.218 0.827742
## speciesD:factor(doy)237          -2.978e-01  8.314e-01  -0.358 0.720210
## speciesE:factor(doy)237           1.265e+00  1.199e+00   1.055 0.291202
## trtflood:factor(doy)194           1.676e-01  2.598e-01   0.645 0.518822
## trtflood:factor(doy)237          -2.228e+00  9.491e-01  -2.347 0.018906 *
## speciesB:trtflood:factor(doy)194  5.764e-02  5.194e-01   0.111 0.911626
## speciesC:trtflood:factor(doy)194 -1.499e-01  3.693e-01  -0.406 0.684691
## speciesD:trtflood:factor(doy)194 -5.262e-02  3.224e-01  -0.163 0.870349
## speciesE:trtflood:factor(doy)194  3.163e-01  4.710e-01   0.672 0.501862
## speciesB:trtflood:factor(doy)237  4.189e-01  1.939e+00   0.216 0.828968
## speciesC:trtflood:factor(doy)237  3.146e-01  1.349e+00   0.233 0.815607
## speciesD:trtflood:factor(doy)237  1.114e+00  1.176e+00   0.947 0.343525
## speciesE:trtflood:factor(doy)237 -2.449e+00  1.695e+00  -1.444 0.148625
## (Intercept).1                    -5.935e+00  2.492e-01 -23.814  < 2e-16 ***
## doy.1                             2.856e-02  1.201e-03  23.775  < 2e-16 ***
## speciesB.1                        5.797e-01  1.105e-01   5.248 1.53e-07 ***
## speciesC.1                        1.026e-02  1.129e-01   0.091 0.927546
## speciesD.1                       -3.148e-01  1.117e-01  -2.819 0.004813 **
## speciesE.1                        3.001e-01  1.195e-01   2.511 0.012040 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##
## Deviance explained = 98.4%
## -REML =  512.5  Scale est. = 1         n = 401
```

```
anova(m3)
```

ANOVA

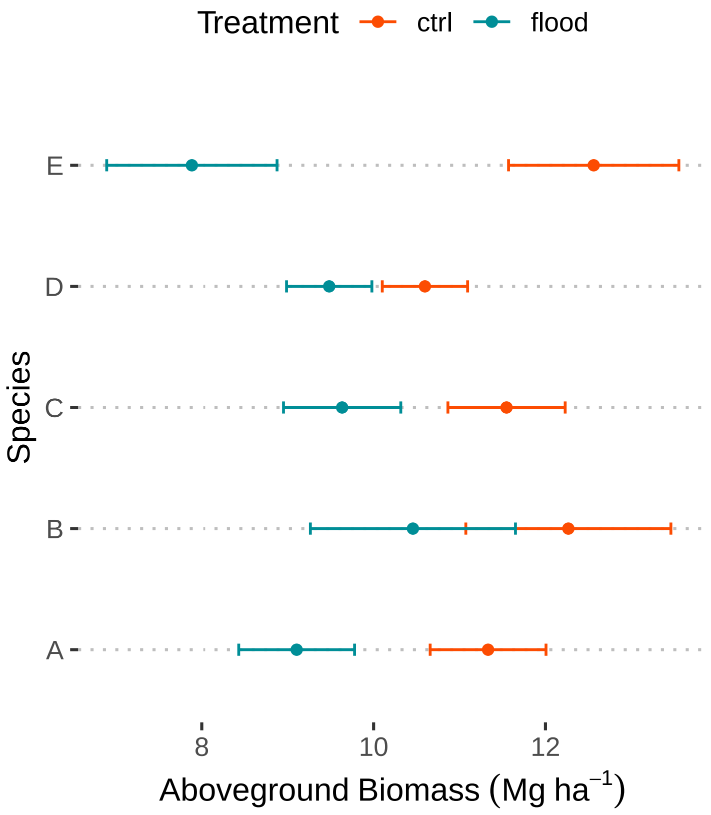
```
##
## Family: gaulss
## Link function: identity logb
##
## Formula:
## agb_g ~ species * trt * factor(doy)
## ~doy + species
##
## Parametric Terms:
##                         df  Chi.sq  p-value
## species                  4   8.404  0.07785
## trt                      1   0.000  1.00000
## factor(doy)              2 364.014  < 2e-16
## species:trt              4   0.000  1.00000
## species:factor(doy)      8  23.406  0.00288
## trt:factor(doy)          2   6.044  0.04870
## species:trt:factor(doy)  8   6.547  0.58619
## doy.1                    1 565.269  < 2e-16
## species.1                4  73.347 4.46e-15
```

```r
means <- emmeans::emmeans(m3,
                          ~trt:species, at = list("doy" = c(237)))

mean_comparisons <- multcomp::cld(means,
                                  level = 0.05,
                                  adjust = "none",
                                  decreasing = TRUE,
                                  Letters = letters) %>%
  mutate(.group = trimws(.group))

mean_comparisons %>%
  transmute(species, trt,SE=round(SE,2),
            emmean = round(emmean, 2), group = .group,
            lower.CL = round(lower.CL,2), upper.CL = round(upper.CL,2))
```

Means, Mean comparisons

```
##    species   trt   SE emmean group lower.CL upper.CL
## 10       E  ctrl 0.99  12.56     a    12.50    12.62
## 2        B  ctrl 1.19  12.27    ab    12.19    12.34
## 8        C  ctrl 0.68  11.55     a    11.50    11.59
## 6        A  ctrl 0.67  11.33    ab    11.29    11.37
## 4        D  ctrl 0.50  10.60   abc    10.57    10.63
## 7        B flood 1.19  10.46  abcd    10.38    10.53
## 1        C flood 0.68   9.63   bcd     9.59     9.68
## 5        D flood 0.50   9.48    cd     9.45     9.51
## 3        A flood 0.67   9.10    cd     9.06     9.15
## 9        E flood 0.99   7.88     d     7.82     7.95
```

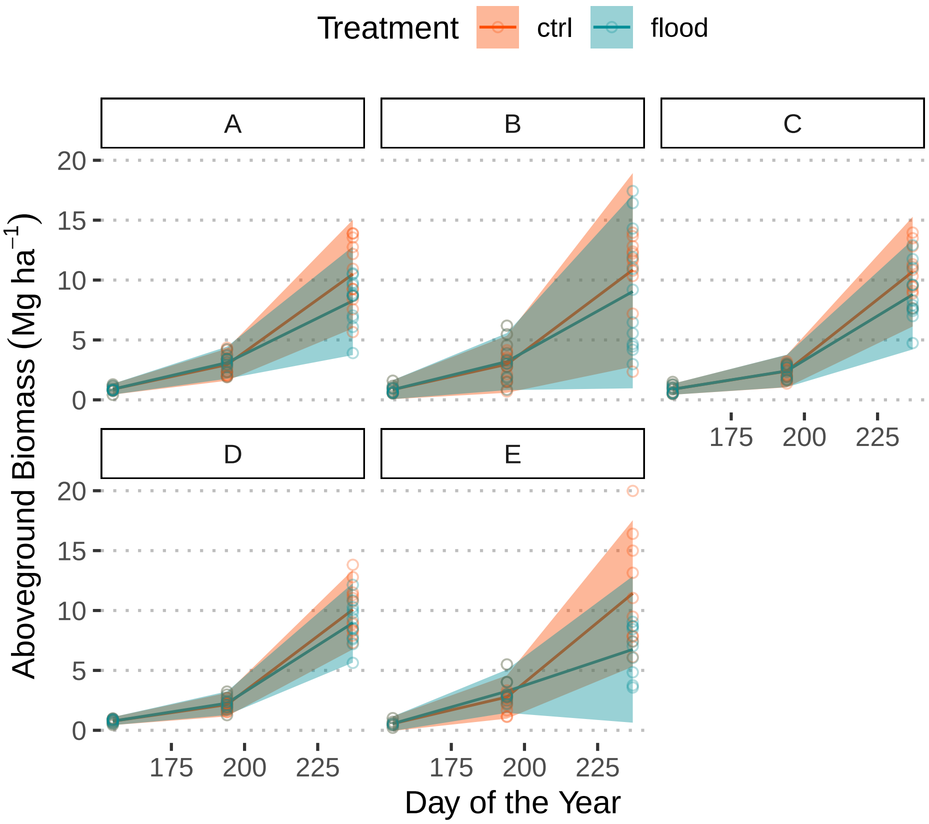
```r
ggplot(mean_comparisons, aes(emmean, species))+
  geom_errorbarh(aes(xmin = emmean-SE, xmax = emmean + SE, color = trt), height = 0.1)+
  geom_point(aes(color = trt) , position = "dodge")+
  scale_fill_manual(values = c("#FC4C02", "#008E97"))+
  scale_color_manual(values = c("#FC4C02", "#008E97"))+
  theme_pubclean()+
  labs(x = expression(Aboveground~Biomass~(Mg~ha^{-1})),
       y = "Species",
       color = "Treatment",
       fill = "Treatment")+
  theme(aspect.ratio = 1)
```



```
plot_data <- expand.grid(doy = unique(dd_finalproj$doy),
                         species = c("A", "B", "C", "D", "E"),
                         trt = c("ctrl", "flood"))

plot_data <- bind_cols(plot_data,
                       predict(m3, newdata = plot_data, type = "link"))
```

```
## New names:
## * ‘‘ -> ‘...4‘
## * ‘‘ -> ‘...5‘
```

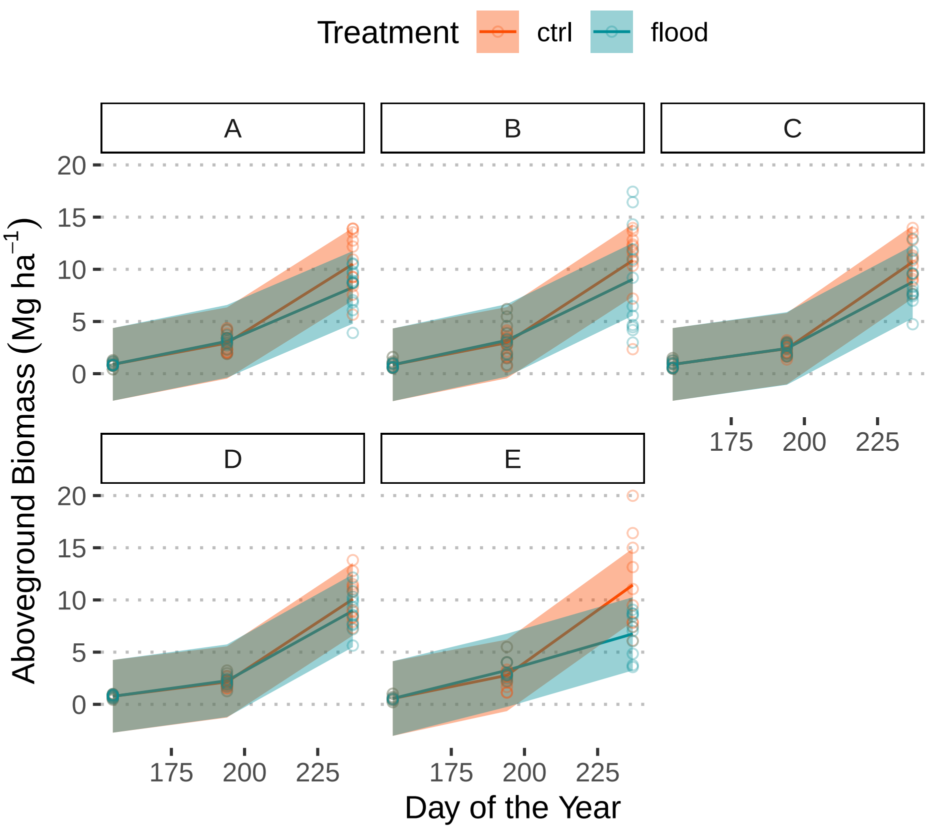
```
plot_data <- plot_data %>%
  rename(fitted = `...4`, sd = `...5`) %>%
  mutate(sd = .01+exp(sd))

plot_data %>%
  ggplot(aes(doy, fitted))+
  geom_line(aes(color = trt))+
  geom_ribbon(aes(ymin = fitted-sd*1.96, ymax = fitted+sd*1.96, fill = trt), alpha = .4)+
  geom_point(aes(y = agb_g, color = trt), data = dd_finalproj, shape = 21, alpha = .3)+
  facet_wrap(~species)+
  scale_fill_manual(values = c("#FC4C02", "#008E97"))+
  scale_color_manual(values = c("#FC4C02", "#008E97"))+
  labs(y = expression(Aboveground~Biomass~(Mg~ha^{-1})),
       x= "Day of the Year",
       color = "Treatment",
       fill = "Treatment")+
  theme_pubclean()+
  theme(aspect.ratio = 1,
        strip.background = element_rect(fill = NA, color = "black"))
```



Bonus: What if we hadn't changed our model?

This is just a demonstration of what happens when we design a model that does not describe our data generating process well. This does not necessarily have to go in your project.

```
plot_data <- expand.grid(doy = unique(dd_finalproj$doy),
                         species = c("A", "B", "C", "D", "E"),
                         trt = c("ctrl", "flood"))

plot_data <- bind_cols(plot_data,
                       predict(m1, newdata = plot_data, interval = "prediction"))

plot_data %>%
  ggplot(aes(doy, fit))+
  geom_line(aes(color = trt))+
  geom_ribbon(aes(ymin = lwr, ymax = upr, fill = trt), alpha = .4)+
  geom_point(aes(y = agb_g, color = trt), data = dd_finalproj, shape = 21, alpha = .3)+
  facet_wrap(~species)+
  scale_fill_manual(values = c("#FC4C02", "#008E97"))+
  scale_color_manual(values = c("#FC4C02", "#008E97"))+
  labs(y = expression(Aboveground~Biomass~(Mg~ha^{-1})),
       x= "Day of the Year",
       color = "Treatment",
       fill = "Treatment")+
  theme_pubclean()+
  theme(aspect.ratio = 1,
        strip.background = element_rect(fill = NA, color = "black"))
```


Treatment
ctrl
flood
A
B
C
D
E
Aboveground Biomass (Mg ha^{-1})
Day of the Year