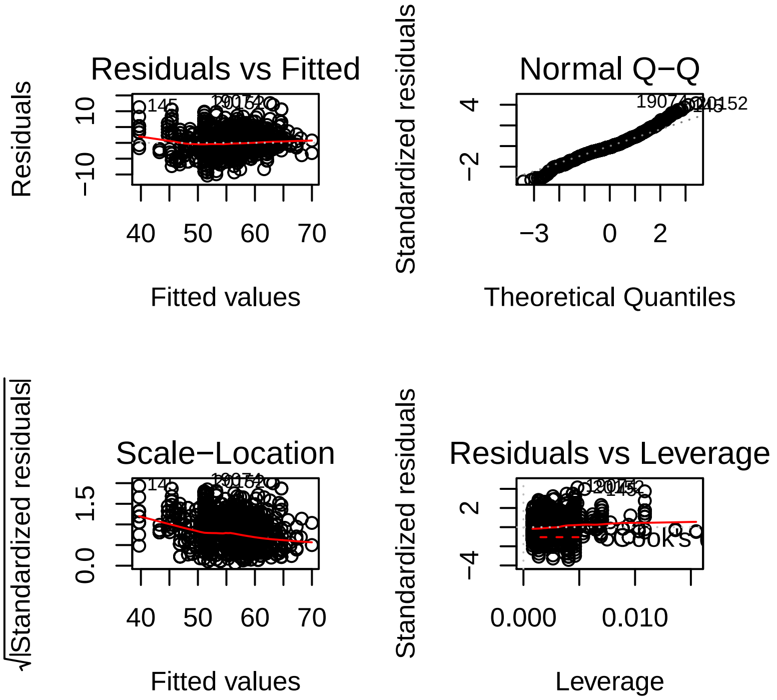
Imputations and data cleaning Lathyrus 2006-2017

I used information on buds sizes to impute FFD. The day when the bud was observed was also included in the model, as well as the interaction, to account for the fact that plants might develop faster/slower in the beginning/end of the season. First, I fitted a model with all years (but note that 2016 and 2017 had no information on bud sizes at all).

```
##
## Call:
## lm(formula = FFD_julian ~ bud_size * day, data = subset_model)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3575  -1.6660  -0.2748   1.6522  12.5720
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.97398    3.11173   5.776 9.18e-09 ***
## bud_size     -26.62119    1.43329 -18.574  < 2e-16 ***
## day            0.34093    0.02466  13.825  < 2e-16 ***
## bud_size:day   0.18339    0.01106  16.584  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.034 on 1588 degrees of freedom
## Multiple R-squared:  0.6923, Adjusted R-squared:  0.6917
## F-statistic:  1191 on 3 and 1588 DF,  p-value: < 2.2e-16
```



Then I fitted a different model for each year

```
model_FFD06<-lm(FFD_julian~bud_size+day,subset(subset_model,year==2006))
model_FFD07<-lm(FFD_julian~bud_size+day,subset(subset_model,year==2007))
model_FFD08<-lm(FFD_julian~bud_size+day,subset(subset_model,year==2008))
model_FFD09<-lm(FFD_julian~bud_size+day,subset(subset_model,year==2009))
model_FFD10<-lm(FFD_julian~bud_size*day,subset(subset_model,year==2010))
model_FFD11<-lm(FFD_julian~bud_size*day,subset(subset_model,year==2011))
model_FFD12<-lm(FFD_julian~bud_size*day,subset(subset_model,year==2012))
model_FFD13<-lm(FFD_julian~bud_size*day,subset(subset_model,year==2013))
model_FFD14<-lm(FFD_julian~bud_size*day,subset(subset_model,year==2014))
model_FFD15<-lm(FFD_julian~bud_size+day,subset(subset_model,year==2015))

summary(model_FFD06)
```

```
##
## Call:
## lm(formula = FFD_julian ~ bud_size + day, data = subset(subset_model,
##     year == 2006))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2636 -2.6960  0.1927  2.3040  9.1184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -65.8096    12.0750  -5.450 4.03e-07 ***
## bud_size     -4.2374     0.5649  -7.501 3.50e-11 ***
## day           1.0186     0.0927  10.988  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.406 on 94 degrees of freedom
## Multiple R-squared:  0.5996, Adjusted R-squared:  0.5911
## F-statistic: 70.39 on 2 and 94 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD07)
```

```
##
## Call:
## lm(formula = FFD_julian ~ bud_size + day, data = subset(subset_model,
##     year == 2007))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2334 -2.4968  0.1884  2.4719  9.1884
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13.96526    5.10969  -2.733  0.00713 **
## bud_size     -4.24567    0.39542 -10.737  < 2e-16 ***
## day           0.61415    0.04479  13.713  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.519 on 133 degrees of freedom
## Multiple R-squared:  0.6101, Adjusted R-squared:  0.6042 
## F-statistic: 104.1 on 2 and 133 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD08)
```

```
## 
## Call:
## lm(formula = FFD_julian ~ bud_size + day, data = subset(subset_model, 
##     year == 2008))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -2.6212 -0.8579 -0.1662  1.1421  4.2782 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -59.71841    4.11840  -14.50   <2e-16 ***
## bud_size     -3.86385    0.30638  -12.61   <2e-16 ***
## day           0.95053    0.03273   29.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1.496 on 79 degrees of freedom
## Multiple R-squared:  0.9222, Adjusted R-squared:  0.9202 
## F-statistic:   468 on 2 and 79 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD09)
```

```
## 
## Call:
## lm(formula = FFD_julian ~ bud_size + day, data = subset(subset_model, 
##     year == 2009))
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.8136 -1.3973 -0.6701  1.2581 10.1864 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -54.56075    6.08939   -8.96 4.81e-15 ***
## bud_size     -3.41625    0.30498  -11.20  < 2e-16 ***
## day           0.90636    0.04794   18.91  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.696 on 121 degrees of freedom
## Multiple R-squared:  0.762,  Adjusted R-squared:  0.7581 
## F-statistic: 193.7 on 2 and 121 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD10)
```

```
## 
## Call:
## lm(formula = FFD_julian ~ bud_size * day, data = subset(subset_model, 
##     year == 2010))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7055 -1.3519 -0.3192  1.2655  7.4597
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   47.58591   11.12482   4.277 2.84e-05 ***
## bud_size     -38.11139    5.03293  -7.572 1.07e-12 ***
## day            0.12104    0.08568   1.413    0.159
## bud_size:day   0.26737    0.03762   7.107 1.72e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.221 on 215 degrees of freedom
## Multiple R-squared:  0.4914, Adjusted R-squared:  0.4843
## F-statistic: 69.25 on 3 and 215 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD11)
```

```
##
## Call:
## lm(formula = FFD_julian ~ bud_size * day, data = subset(subset_model,
##     year == 2011))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5094 -1.7585 -0.3598  1.4642 12.2415
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.40431    9.43885  -0.573    0.567
## bud_size     -17.58628    3.79647  -4.632 5.82e-06 ***
## day            0.50497    0.07584   6.659 1.74e-10 ***
## bud_size:day   0.12020    0.02977   4.038 7.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.837 on 250 degrees of freedom
## Multiple R-squared:  0.6268, Adjusted R-squared:  0.6223
## F-statistic:   140 on 3 and 250 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD12)
```

```
##
## Call:
## lm(formula = FFD_julian ~ bud_size * day, data = subset(subset_model,
##     year == 2012))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5476 -1.1082  0.0448  1.0528  7.4399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   32.83286    4.49883   7.298 1.40e-12 ***
## bud_size     -28.14901    2.51719 -11.183  < 2e-16 ***
## day            0.23258    0.03602   6.457 2.87e-10 ***
## bud_size:day   0.19421    0.01917  10.129  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.511 on 435 degrees of freedom
## Multiple R-squared:  0.6248, Adjusted R-squared:  0.6222
## F-statistic: 241.4 on 3 and 435 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD13)
```

```
##
## Call:
## lm(formula = FFD_julian ~ bud_size * day, data = subset(subset_model,
##     year == 2013))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4899 -1.6008 -0.4331  1.8437  5.5101
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.68489   23.50311   1.688   0.0954 .
## bud_size     -19.30547    7.82932  -2.466   0.0159 *
## day            0.15396    0.17564   0.877   0.3835
## bud_size:day   0.13971    0.05789   2.414   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.03 on 77 degrees of freedom
## Multiple R-squared:  0.3706, Adjusted R-squared:  0.346
## F-statistic: 15.11 on 3 and 77 DF,  p-value: 7.985e-08
```

```
summary(model_FFD14)
```

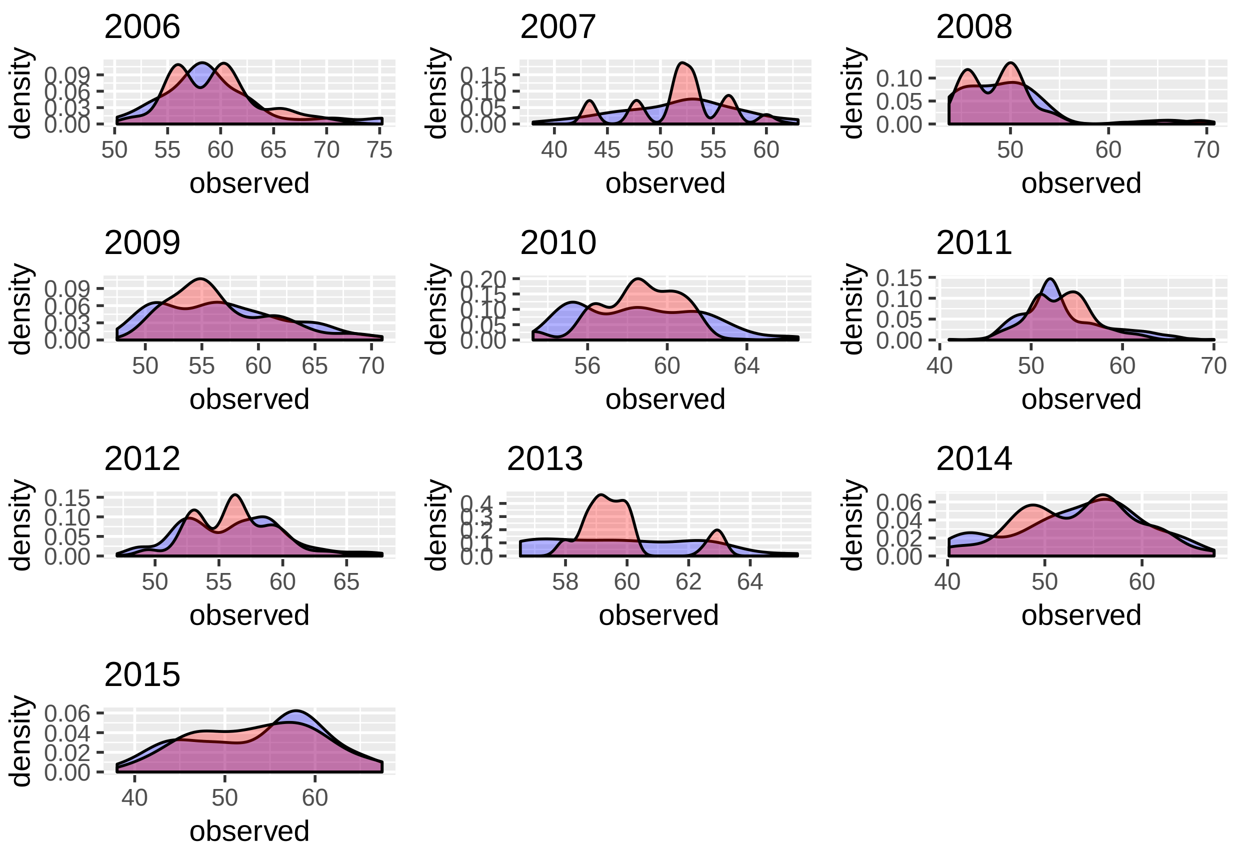
```
##
## Call:
## lm(formula = FFD_julian ~ bud_size * day, data = subset(subset_model,
##     year == 2014))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6612 -1.3338  0.3508  1.3690  8.3872
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.67648    9.75090  -2.223 0.028240 *
## bud_size    -17.68200    5.22002  -3.387 0.000977 ***
## day           0.64733    0.07773   8.328  2.4e-13 ***
## bud_size:day  0.10873    0.04069   2.672 0.008674 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 2.737 on 111 degrees of freedom
## Multiple R-squared:  0.8352, Adjusted R-squared:  0.8307
## F-statistic: 187.5 on 3 and 111 DF,  p-value: < 2.2e-16
```

```
summary(model_FFD15)
```

```
##
## Call:
## lm(formula = FFD_julian ~ bud_size + day, data = subset(subset_model,
##     year == 2015))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8412 -2.1725 -0.5807  1.1591  8.6974
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -71.74481    7.54831  -9.505 5.00e-12 ***
## bud_size     -5.13008    0.57019  -8.997 2.39e-11 ***
## day           1.07690    0.06461  16.669  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.692 on 42 degrees of freedom
## Multiple R-squared:  0.8689, Adjusted R-squared:  0.8626
## F-statistic: 139.2 on 2 and 42 DF,  p-value: < 2.2e-16
```

Compare distributions of observed vs predicted values



Results of the imputation

```
nrow(subset(data_imput,FFD_action=="impute"&!is.na(FFD))) #56 cases imputed
```

```
## [1] 56
```

```
nrow(subset(data_imput,FFD_action=="impute"&is.na(FFD))) #46 cases could not be imputed because no info
```

```
## [1] 46
```

```
nrow(subset(data_imput,is.na(FFD)&data==1)) #46 cases where FFD is NA = cases that could not be imputed
```

```
## [1] 46
```

Predict shoot volume with number of flowers?

```
nrow(subset(data_imput,data==1&is.na(shoot_vol))) #37 pls with no shoot_vol
```

```
## [1] 37
```

```
nrow(subset(data_imput,data==1&is.na(cum_n_fl))) #24 pls with no cum_n_fl
```

```
## [1] 24
```

```
nrow(subset(data_imput,data==1&is.na(cum_n_fl)&is.na(shoot_vol)))
```

```
## [1] 3
```

```
#3 pls with no shoot_vol and no cum_n_fl --> mark as missing data
nrow(subset(data_imput,is.na(cum_n_fl)&data==1&!is.na(shoot_vol)))
```

```
## [1] 21
```

```
#We can predict shoot_vol from cum_n_fl in 24-3=21 pls
```