# 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
## -10.7355  -1.8024  -0.2656   1.7064  12.6290
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  100.40237    3.14664   31.91   <2e-16 ***
## bud_size     -26.90853    1.44937  -18.57   <2e-16 ***
## day            0.32035    0.02494   12.85   <2e-16 ***
## bud_size:day   0.18610    0.01118   16.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.068 on 1588 degrees of freedom
## Multiple R-squared:  0.6792, Adjusted R-squared:  0.6786
## F-statistic:  1121 on 3 and 1588 DF,  p-value: < 2.2e-16
```
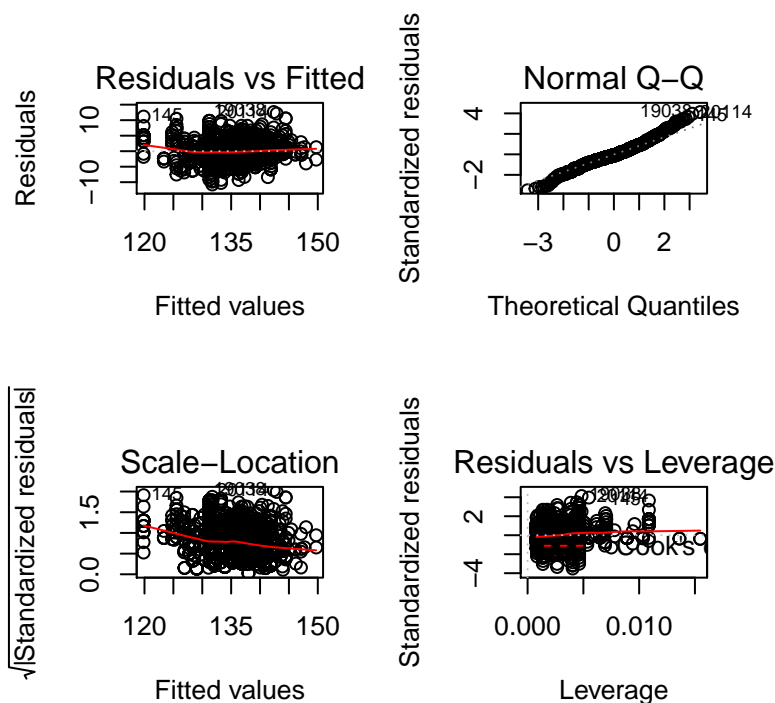
```r
sjp.int(model_FFD, type = "eff",mdrt.values = "quart",swap.pred=T) #Plot of interaction effect
```

```
## `sjp.int()` will become deprecated in the future. Please use `plot_model()` instead.
```


Interaction effect of day and bud_size on FFD_julian
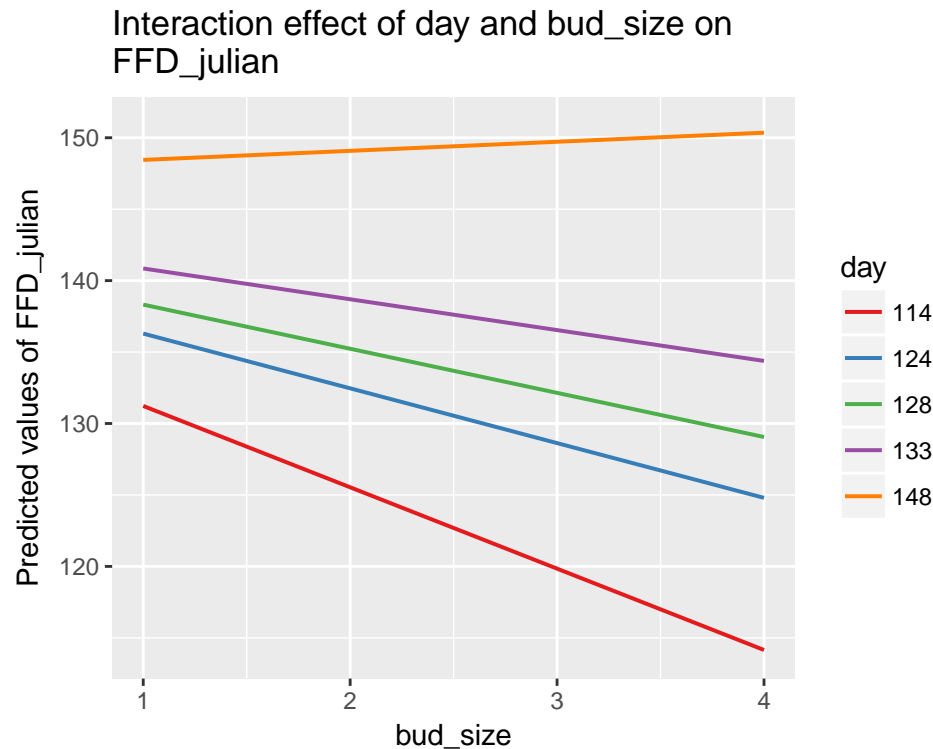
Then I fitted a different model for each year

```r
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)   13.9578    12.0750    1.156     0.251
## bud_size       -4.2374     0.5649   -7.501   3.5e-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) 66.04446    5.10969   12.93   <2e-16 ***
## bud_size    -4.24567    0.39542  -10.74   <2e-16 ***
## day          0.61415    0.04479   13.71   <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) 20.53437    4.11840    4.986 3.57e-06 ***
## bud_size    -3.86385    0.30638  -12.611  < 2e-16 ***
## day          0.95053    0.03273   29.043  < 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) 24.93439    6.08939   4.095 7.67e-05 ***
## bud_size    -3.41625    0.30498 -11.201  < 2e-16 ***
## day          0.90636    0.04794  18.907  < 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)  127.32341   11.12482  11.445  < 2e-16 ***
## 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)  74.57555    9.43885   7.901 8.81e-14 ***
## 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)  113.05508    4.49883  25.130  < 2e-16 ***
## 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)  119.15016   23.50311   5.070 2.67e-06 ***
## 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)   58.03116    9.75090   5.951 3.16e-08 ***
## bud_size     -17.68200    5.22002  -3.387 0.000977 ***
## day            0.64733    0.07773   8.328 2.40e-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)  8.20519    7.54831   1.087    0.283
## 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
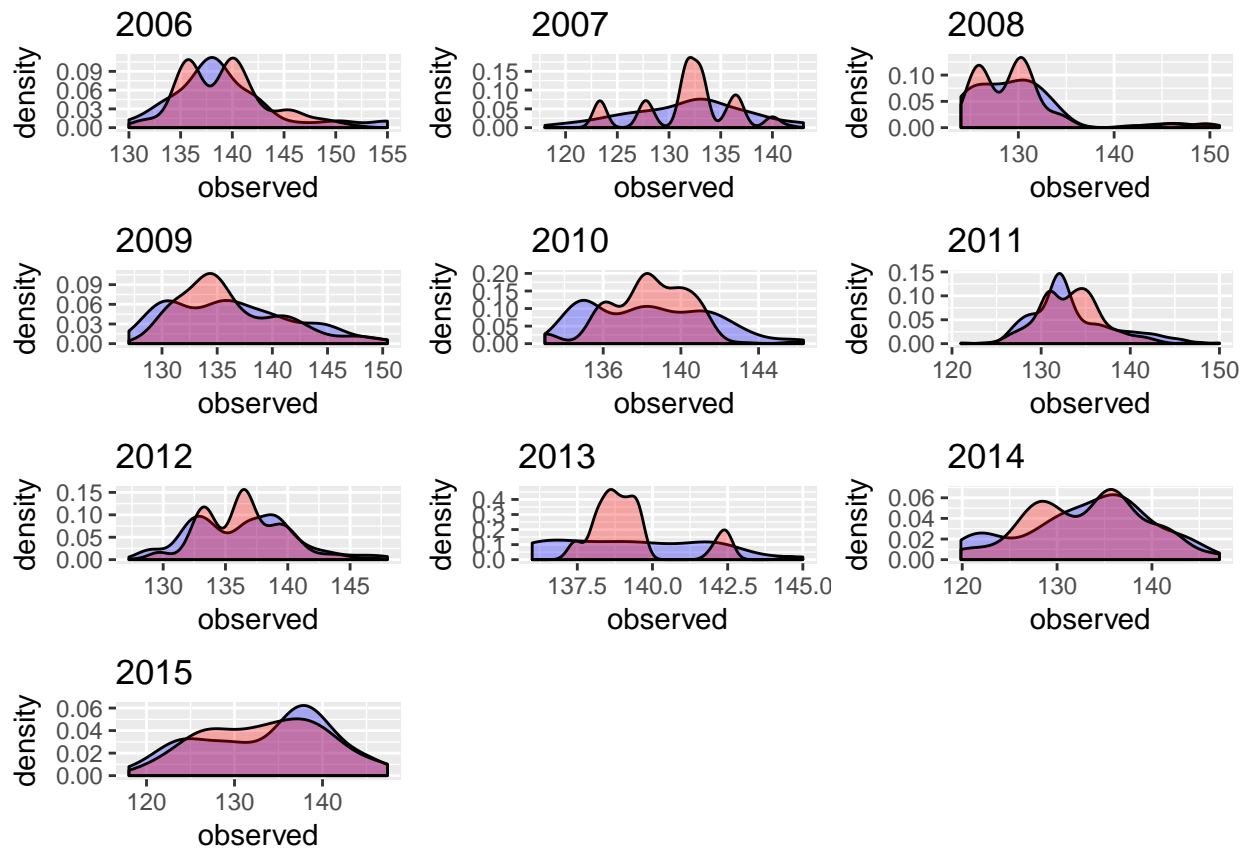
Results of the imputation

```r
nrow(subset(data_imput,FFD_action=="impute"&!is.na(FFD))) #168 cases imputed
```

```
## [1] 168
```

```r
nrow(subset(data_imput,FFD_action=="impute"&is.na(FFD)) )
```

```
## [1] 11
```

```r
#11 cases could not be imputed because no info on bud sizes - What to do?
```

Predict shoot volume with number of flowers?

```r
nrow(subset(data_imput,data==1&is.na(shoot_vol))) #40 pls with no shoot_vol
```

```
## [1] 40
```

```r
nrow(subset(data_imput,data==1&is.na(cum_n_fl))) #31 pls with no cum_n_fl
```

```
## [1] 31
```

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

```
## [1] 4
```

```r
#4 pls with no shoot_vol and no cum_n_fl
#2 of them where FFD was imputed (use?), 2 also with no FFD (remove?)
```
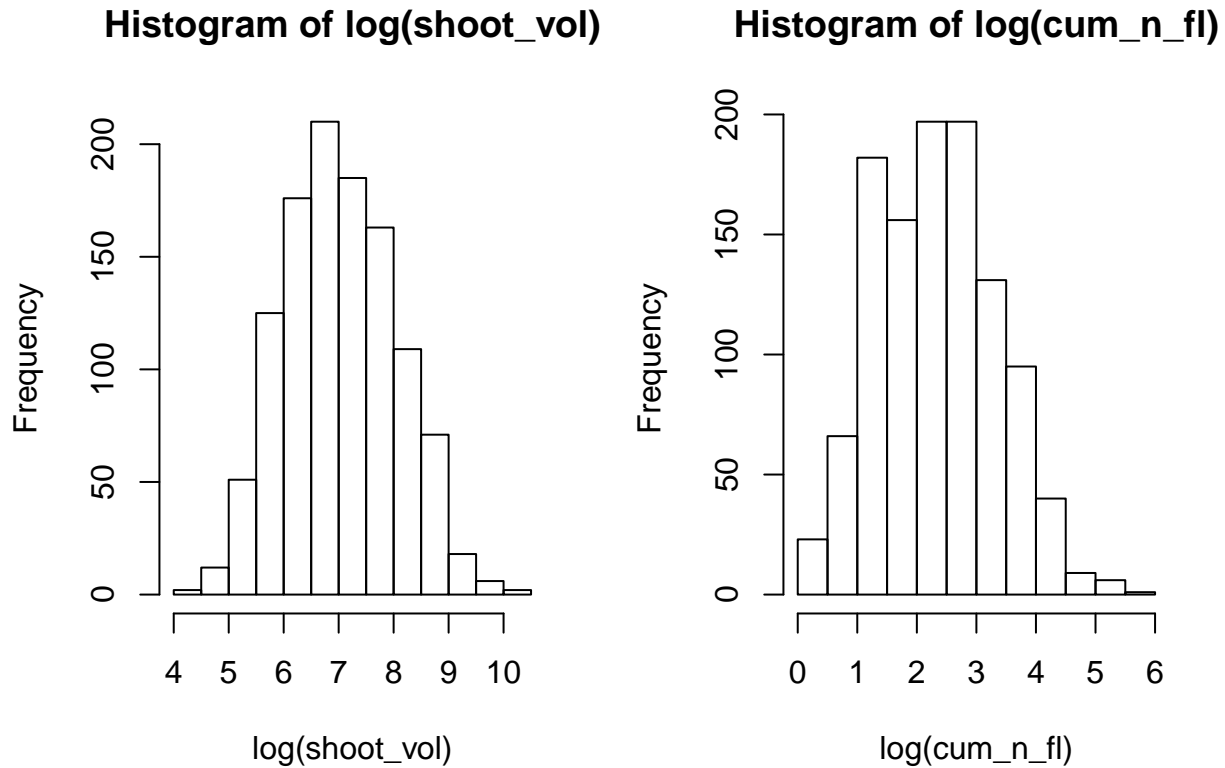
We can predict shoot_vol from cum_n_fl in 40-4=36 pls

```r
nrow(subset(data_imput,data==1&shoot_vol==0)) #1 pl with shoot_vol=0, impute?
```

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

Histograms

Histogram of log(shoot_vol)

Histogram of log(cum_n_fl)

Linear model

```r
model_vol<-lm(log(shoot_vol)~log(cum_n_fl),subset(data_imput,data==1&shoot_vol>0))
summary(model_vol)
```
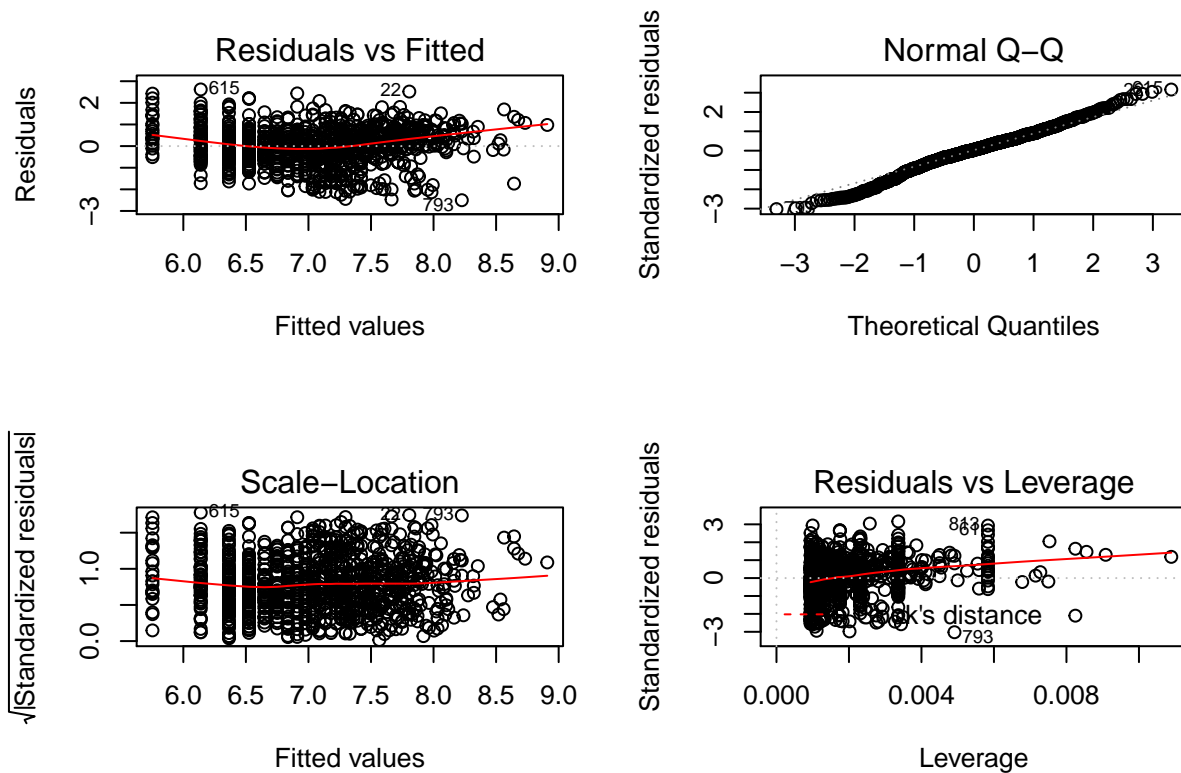
```
##
## Call:
## lm(formula = log(shoot_vol) ~ log(cum_n_fl), data = subset(data_imput,
##     data == 1 & shoot_vol > 0))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50298 -0.45651  0.02757  0.50103  2.61984
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.75322    0.06340   90.75   <2e-16 ***
## log(cum_n_fl)  0.55658    0.02485   22.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```
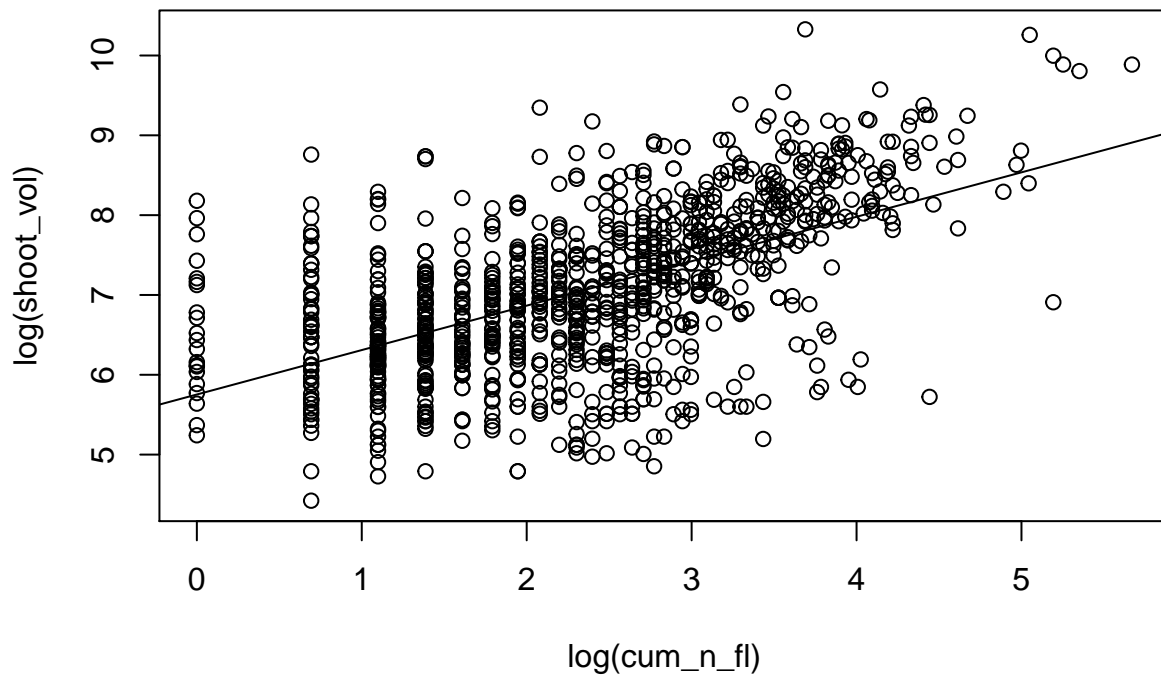
```
## Residual standard error: 0.83 on 1064 degrees of freedom
##   (27 observations deleted due to missingness)
## Multiple R-squared:  0.3205, Adjusted R-squared:  0.3198
## F-statistic: 501.8 on 1 and 1064 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(model_vol)
```



```r
par(mfrow=c(1,1))
with(subset(data_imput,data==1&shoot_vol>0),plot(log(shoot_vol)~log(cum_n_fl)))
abline(model_vol)
```

Negative binomial model

```
summary(model_vol_nb)
```

```
##
## Call:
## glm.nb(formula = shoot_vol ~ cum_n_fl, data = subset(data_imput,
##     data == 1 & shoot_vol > 0), init.theta = 1.616244962, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6484  -0.9347  -0.3403   0.2936   4.6622
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.905624   0.030256  228.24   <2e-16 ***
## cum_n_fl    0.027906   0.001043   26.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.6162) family taken to be 1)
##
##     Null deviance: 1799.0  on 1065  degrees of freedom
## Residual deviance: 1171.8  on 1064  degrees of freedom
##   (27 observations deleted due to missingness)
## AIC: 17776
##
```

```
## Number of Fisher Scoring iterations: 1
##
##
##              Theta:  1.6162
##          Std. Err.:  0.0642
##
##  2 x log-likelihood:  -17769.5950
```

```
NagelkerkeR2(model_vol_nb)
```

```
## $N
## [1] 1066
##
## $R2
## [1] 0.5457461
```

```
par(mfrow=c(2,2))
plot(model_vol_nb)
```