Assignment2-446-Meng Gao

Introduction

In this report, we analyze two datasets. The Buttercup dataset looks at the efficacy of different herbicides and management of pasture on the weed, giant buttercup (Ranunculus Acris), growth in dairy pastures, and we will be fitting linear mixed and a generalized linear mixed models. We will fit a generalized additive model on the Covid cases dataset.

QUESTION 1 FITTING A LINEAR MIXED MODEL

We start by importing and cleansing the data and drawing some graphs to show the important features.

```
buttercup <- read.csv("/Charlotte/shepherding/446GLM/assignment2/Buttercup\
 Data.csv")
str(buttercup)
```

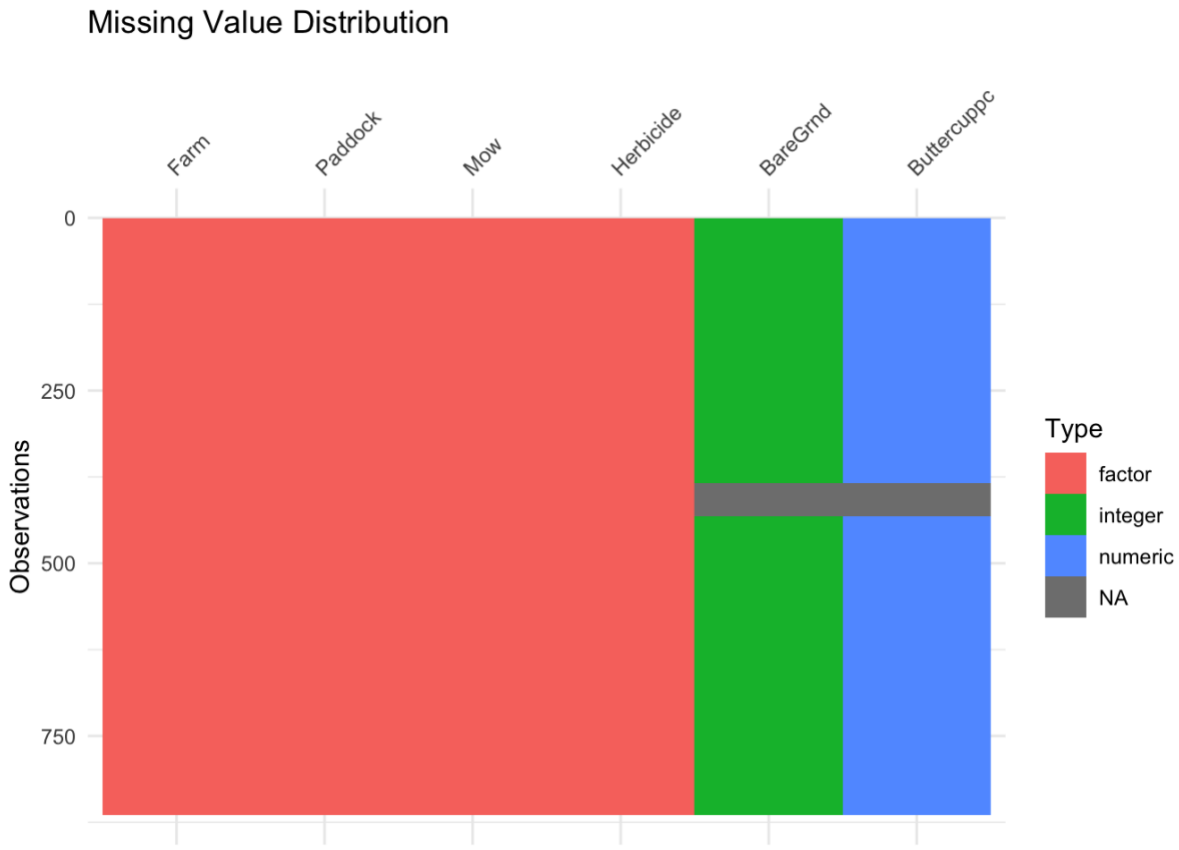
```
## 'data.frame':    864 obs. of  6 variables:
##  $ Farm       : Factor w/ 9 levels "Farm A","Farm B",..: 2 2 2 2 2 2 2 2
 2 2 ...
##  $ Paddock    : Factor w/ 2 levels "Dry","Wet": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Mow        : Factor w/ 2 levels "Mow","No Mow": 1 1 1 1 1 1 1 1 1 1
 ...
##  $ Herbicide  : Factor w/ 8 levels "Aminopyralid",..: 6 5 3 6 4 7 4 6 7
8 ...
##  $ Buttercuppc: num  0.1 0.55 0.05 0 0.05 1.05 0 0 0.75 0.25 ...
##  $ BareGrnd   : int  110 125 125 188 175 150 188 175 150 175 ...
```

```
summary(buttercup)
```

```
##       Farm     Paddock       Mow                         Herbicide
##  Farm A : 96   Dry:432   Mow   :432   Aminopyralid          :108
##  Farm B : 96   Wet:432   No Mow:432   Aminopyralid+triclopyr:108
##  Farm C : 96                          Flumetsulam           :108
##  Farm D : 96                          MCPA                  :108
##  Farm E : 96                          MCPB                  :108
##  Farm F : 96                          MCPB+bentazone        :108
##  (Other):288                          (Other)               :216
##   Buttercuppc        BareGrnd
##  Min.   : 0.000   Min.   :  0.00
```

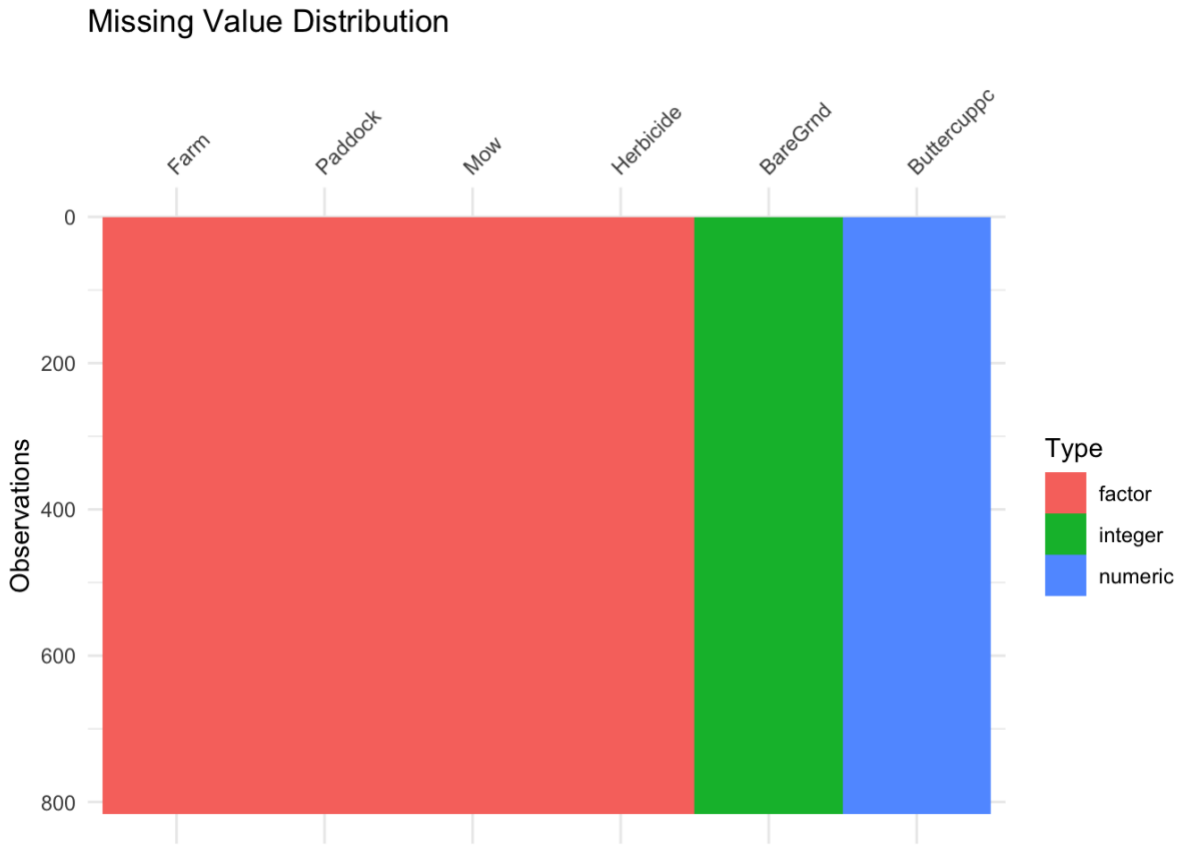
```
##  1st Qu.: 0.100   1st Qu.: 15.00
##  Median : 1.025   Median : 55.00
##  Mean   : 4.204   Mean   : 81.23
##  3rd Qu.: 4.250   3rd Qu.:115.00
##  Max.   :65.000   Max.   :525.00
##  NA's   :48       NA's   :48
```

Looking at the summary can see that the dataset is almost balanced (counts are equal in each of the variety levels and block levels except the "Other" variable). Let's check if there is any missing values and remove them if there is
any.

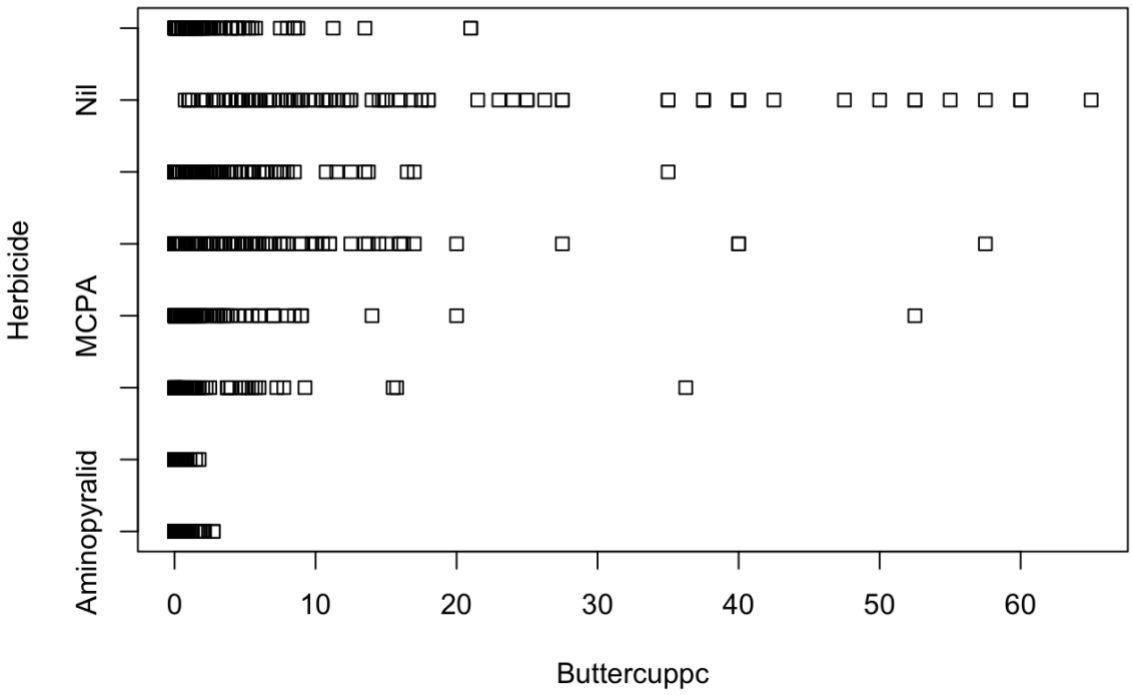


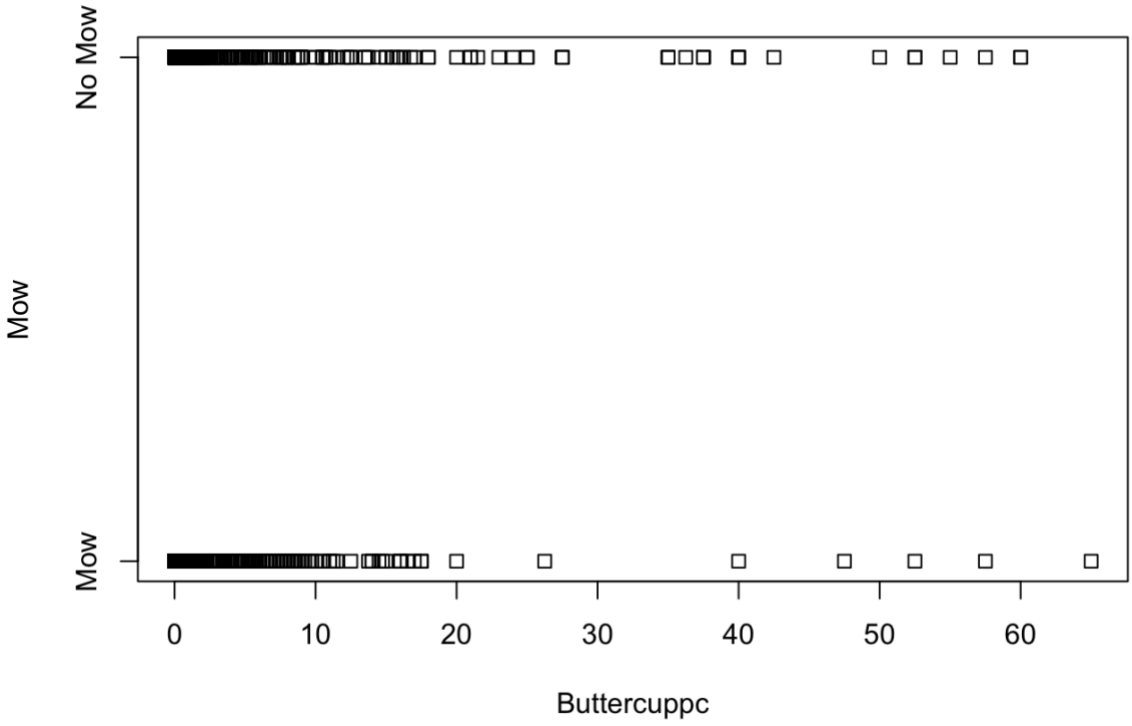
As can be seen in the missing value distrubition plot, there appears to be a missing pattern.

We check out the missing value distribution again after omitting the missing values.



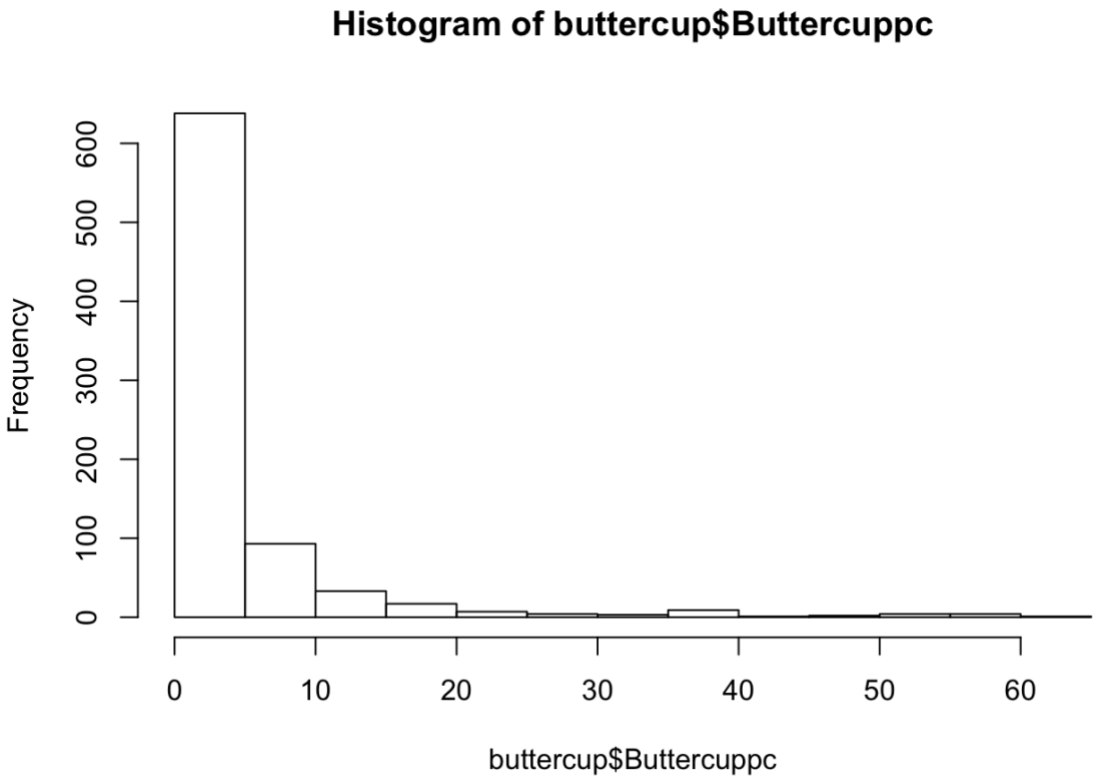
We can also embed one dimensional scatter plots of the given data.





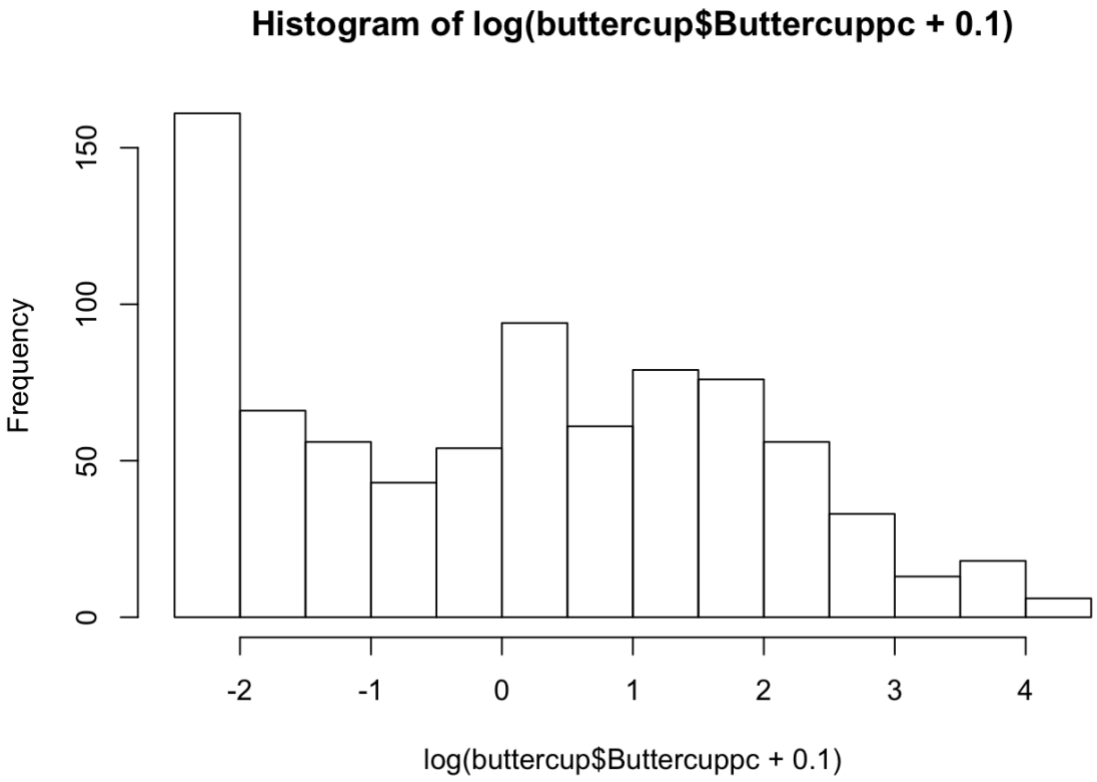
The scatterplots show that it does not make much difference if the pasture is mown or not, all herbicides work for buttercup but the affects can be significantly different.

Before we fit any models, we plot a histogram of the response variable (Buttercuppc) to check how it distributes and will need to consider a suitable transformation for it if it is not normal distribution.



We will add a small amount to the zeros and then take logs, this avoids the issue of the log of zero being

undefined.



The distrubution after log transformation looks fine, now we generate a new variable which is the interaction between the Farm and Paddock variables.

Now we fit a simple analysis of variance model to the data including the interaction between Herbicide and Mow and a term for the Farm and Paddock as above.

```
##                Df Sum Sq Mean Sq F value  Pr(>F)
## Herbicide       7 1254.2  179.17 189.658 < 2e-16 ***
## Mow             1    8.1    8.13   8.608 0.00344 **
## FarmPaddock    16  546.6   34.16  36.161 < 2e-16 ***
## Herbicide:Mow   7   16.1    2.30   2.434 0.01798 *
## Residuals     784  740.6    0.94
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary of ANOVA, we can see that there is a significant difference between all the independent terms used for the model, as all the p-values are less than 0.05.

Now we fit a random effects model with the Farm by Paddock variable as the random effect and fixed effects for mowing and herbicides and their interaction in the model.

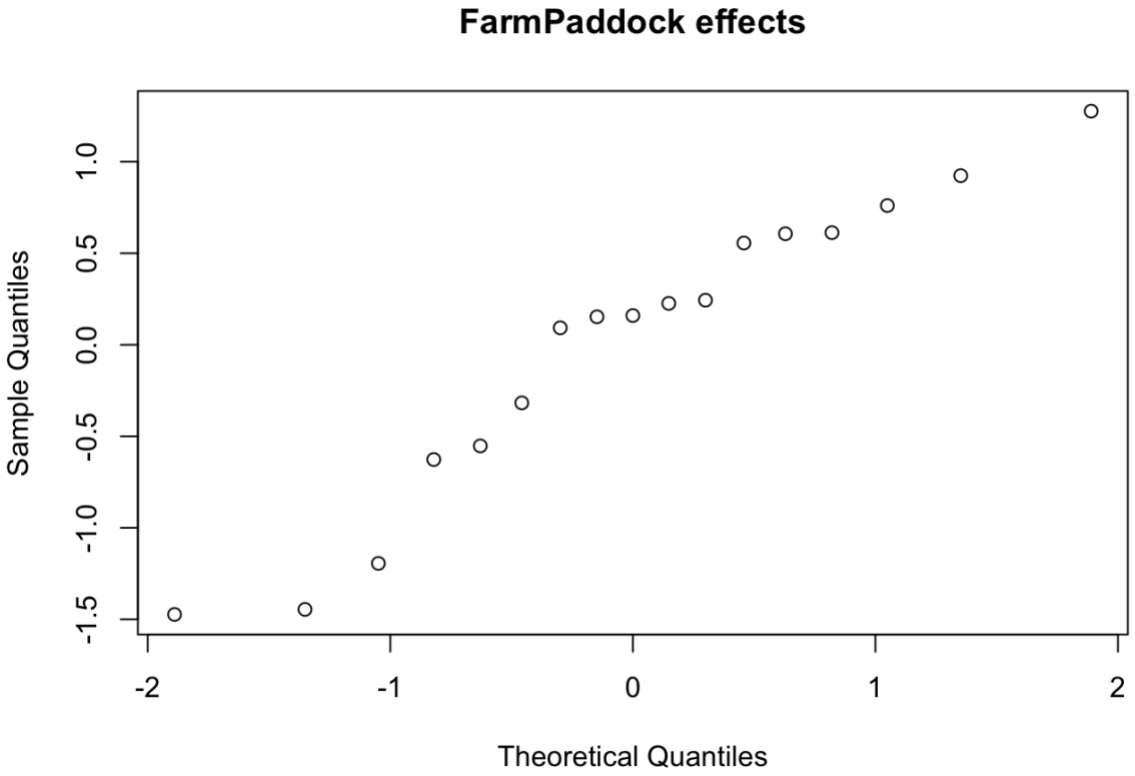
```
## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: log(buttercup$Buttercuppc + 0.1) ~ Herbicide * Mow + (1 | FarmP
addock)
##    Data: buttercup
##
## REML criterion at convergence: 2345.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4369 -0.6747  0.0120  0.6734  3.5450
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  FarmPaddock (Intercept) 0.6920   0.8319
##  Residual                0.9447   0.9719
## Number of obs: 816, groups:  FarmPaddock, 17
##
## Fixed effects:
##                                          Estimate Std. Error t value
## (Intercept)                              -1.48338    0.24337  -6.095
## HerbicideAminopyralid+triclopyr          -0.41433    0.19247  -2.153
## HerbicideFlumetsulam                      0.97800    0.19247   5.081
## HerbicideMCPA                             1.35347    0.19247   7.032
## HerbicideMCPB                             2.40947    0.19247  12.518
## HerbicideMCPB+bentazone                   1.94594    0.19247  10.110
## HerbicideNil                              3.39159    0.19247  17.621
## HerbicideThifensulfuron-methyl            1.88632    0.19247   9.800
## MowNo Mow                                 0.20817    0.19247   1.082
## HerbicideAminopyralid+triclopyr:MowNo Mow -0.02504    0.27220  -0.092
## HerbicideFlumetsulam:MowNo Mow            -0.16168    0.27220  -0.594
## HerbicideMCPA:MowNo Mow                   -0.11290    0.27220  -0.415
## HerbicideMCPB:MowNo Mow                    0.24233    0.27220   0.890
## HerbicideMCPB+bentazone:MowNo Mow         -0.12554    0.27220  -0.461
## HerbicideNil:MowNo Mow                     0.56296    0.27220   2.068
## HerbicideThifensulfuron-methyl:MowNo Mow  -0.44822    0.27220  -1.647
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
```

```
##     vcov(x)        if you need it
```

The Farm by Paddock term explains a lot about the total variability. By comparing the Farm by Paddock effect to the residual in the ANOVA model in 3, We can know that the random effects the sums of squares accounted for by Farm by Paddock effect is 546.6(21.30%), while the residual sum of squares accounts for 740.6(28.87%). Looking at the random effects the variance accounted for by Farm by Paddock is 0.6920 (42.28%), while the residual variance(variance un-accounted for by the model) is 0.9447(55.72%).

We would also be checking normality of the random effects.



Looking at the Q-Q plot we can see approximate normality in FarmPaddock.

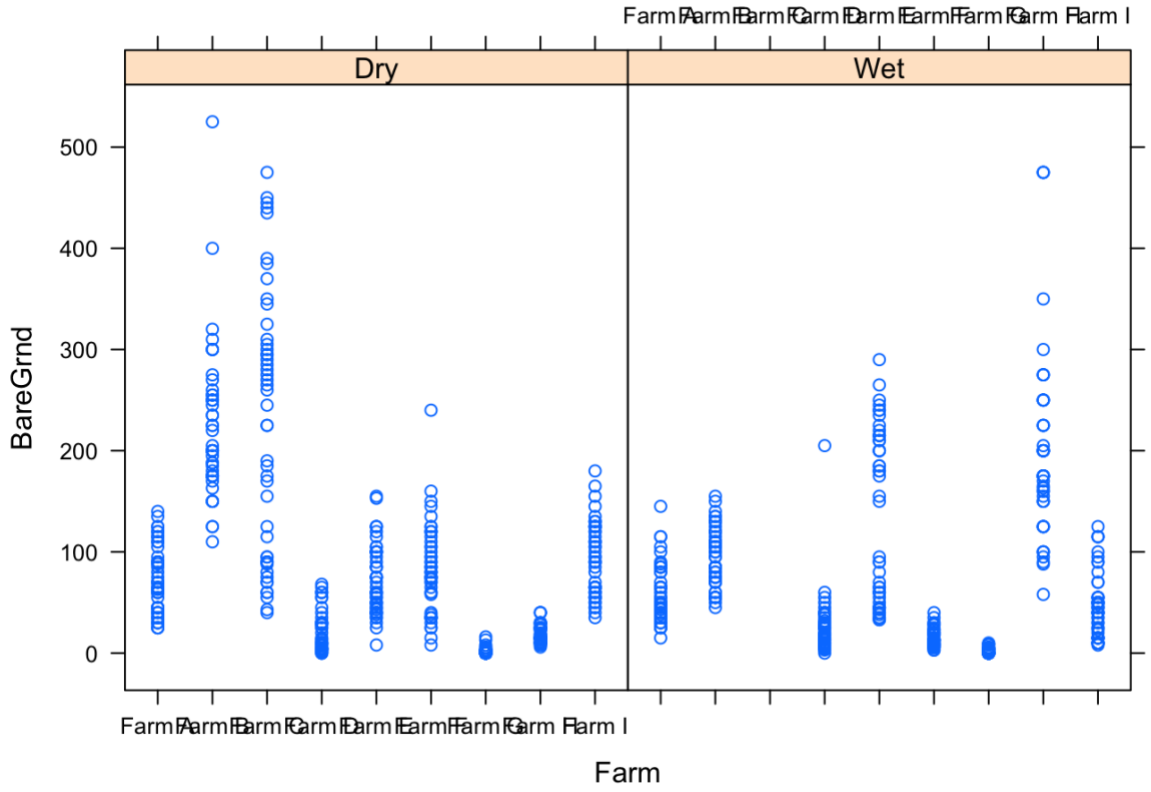
Ducussion of Question1

We cleanse the data by deleting the null variables and draw stripcharts to explore the relationships between Buttercuppc and Herbicide or Mow. We log transform the response variable to make sure it is approximate normal distribution. We then gererate an interaction term of Farm and Paddock and fit a simple analysis of variance model and a random effects model with this included. By looking at the summary oututs of these two models, we find this interaction term explains a useful amount of the variablity.

QUESTION 2 FITTING A GENERALISED LINEAR MIXED MODEL

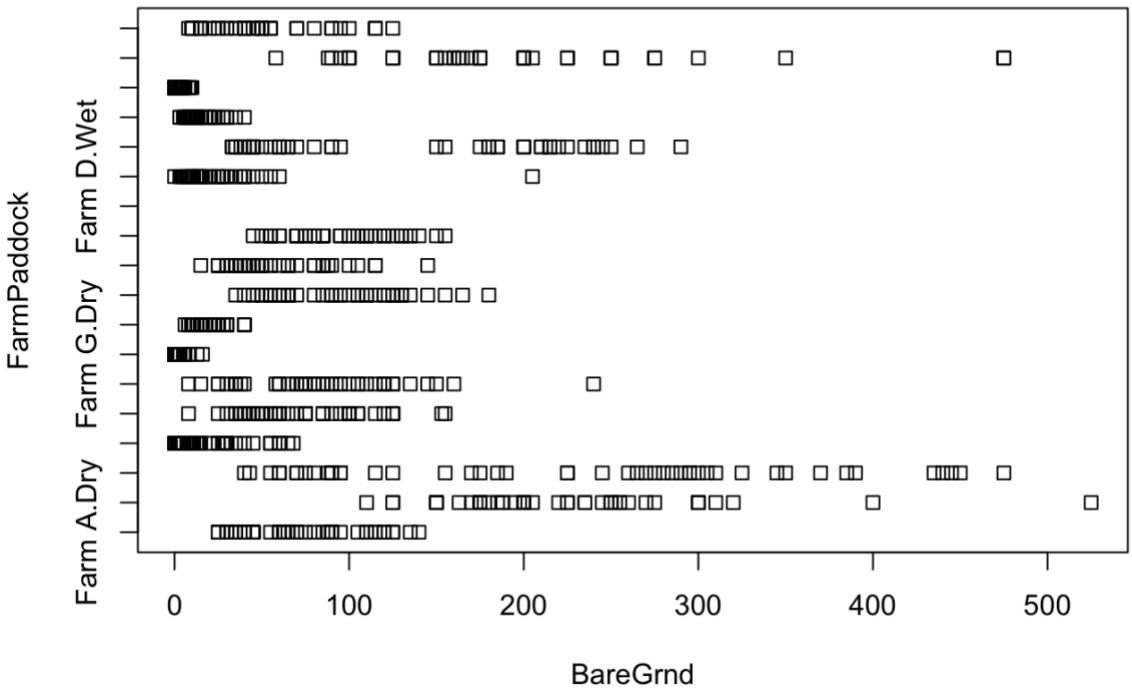
We first examine the relationship between BareGrnd and Paddock individually for Farm, using the xyplot() function. Looking at this plot, there are some farms that have more variability than others, there is also variability for dry and wet farms.

```
## Warning: package 'lattice' was built under R version 3.6.2
```



We now look at the stripchart for the interaction variable. This plot shows that for different levels in the interaction term, BareGrnd is significantly different, it makes sense to include the interaction term in the

model.



We now fit a Generalised Linear model to the data with fixed effects for both Mow and Herbicide and their interaction and a term for the Farm by Paddock interaction. As it says in the question, we can approximate this variable as an unbounded count so we can assume the response variable is poisson distribution.

```
##
## Call:
## glm(formula = buttercup$BareGrnd ~ Herbicide * Mow + FarmPaddock,
##     family = poisson, data = buttercup)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -12.287   -2.702   -0.728    2.089   21.150
##
## Coefficients:
##                                          Estimate Std. Error z value Pr
(>|z|)
## (Intercept)                               4.63550    0.02108 219.941  <
 2e-16
## HerbicideAminopyralid+triclopyr           0.02406    0.01816   1.325
0.18508
```

```
## HerbicideFlumetsulam                  -0.28219    0.01970 -14.326  <
 2e-16

## HerbicideMCPA                         -0.15290    0.01900  -8.046 8.
59e-16

## HerbicideMCPB                         -0.36133    0.02015 -17.928  <
 2e-16

## HerbicideMCPB+bentazone               -0.31731    0.01990 -15.947  <
 2e-16

## HerbicideNil                          -0.30095    0.01980 -15.197  <
 2e-16

## HerbicideThifensulfuron-methyl        -0.30728    0.01984 -15.488  <
 2e-16

## MowNo Mow                             -0.43329    0.02059 -21.041  <
 2e-16

## FarmPaddockFarm B.Dry                  1.13047    0.01967  57.471  <
 2e-16

## FarmPaddockFarm C.Dry                  1.14565    0.01963  58.350  <
 2e-16

## FarmPaddockFarm D.Dry                 -1.21841    0.03580 -34.034  <
 2e-16

## FarmPaddockFarm E.Dry                  0.02712    0.02402   1.129
0.25891

## FarmPaddockFarm F.Dry                  0.14487    0.02336   6.203 5.
55e-10

## FarmPaddockFarm G.Dry                 -3.76766    0.11380 -33.107  <
 2e-16

## FarmPaddockFarm H.Dry                 -1.31491    0.03717 -35.373  <
 2e-16

## FarmPaddockFarm I.Dry                  0.26916    0.02271  11.850  <
 2e-16

## FarmPaddockFarm A.Wet                 -0.21050    0.02556  -8.234  <
 2e-16

## FarmPaddockFarm B.Wet                  0.28358    0.02264  12.523  <
 2e-16

## FarmPaddockFarm D.Wet                 -1.04619    0.03354 -31.190  <
 2e-16

## FarmPaddockFarm E.Wet                  0.62726    0.02118  29.613  <
 2e-16

## FarmPaddockFarm F.Wet                 -1.56342    0.04110 -38.040  <
 2e-16

## FarmPaddockFarm G.Wet                 -3.65977    0.10796 -33.898  <
 2e-16

## FarmPaddockFarm H.Wet                  1.03075    0.01992  51.743  <
 2e-16

## FarmPaddockFarm I.Wet                 -0.40781    0.02706 -15.071  <
 2e-16
```

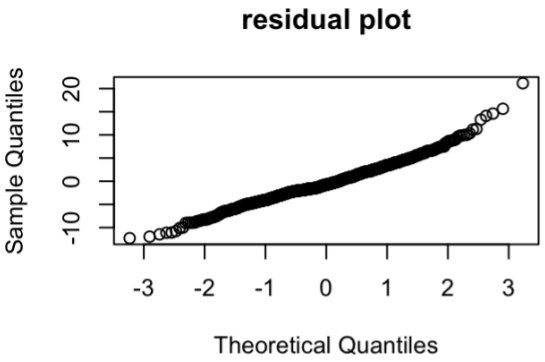
```
## HerbicideAminopyralid+triclopyr:MowNo Mow  0.19439    0.02817   6.900 5.
20e-12
## HerbicideFlumetsulam:MowNo Mow             0.01949    0.03130   0.623
0.53344
## HerbicideMCPA:MowNo Mow                    0.05059    0.03006   1.683
0.09232
## HerbicideMCPB:MowNo Mow                    0.12472    0.03146   3.965 7.
34e-05
## HerbicideMCPB+bentazone:MowNo Mow          0.16294    0.03088   5.277 1.
31e-07
## HerbicideNil:MowNo Mow                    -0.24354    0.03306  -7.367 1.
75e-13
## HerbicideThifensulfuron-methyl:MowNo Mow   0.11809    0.03101   3.808
0.00014
##
## (Intercept)                               ***
## HerbicideAminopyralid+triclopyr
## HerbicideFlumetsulam                      ***
## HerbicideMCPA                             ***
## HerbicideMCPB                             ***
## HerbicideMCPB+bentazone                   ***
## HerbicideNil                              ***
## HerbicideThifensulfuron-methyl            ***
## MowNo Mow                                 ***
## FarmPaddockFarm B.Dry                     ***
## FarmPaddockFarm C.Dry                     ***
## FarmPaddockFarm D.Dry                     ***
## FarmPaddockFarm E.Dry
## FarmPaddockFarm F.Dry                     ***
## FarmPaddockFarm G.Dry                     ***
## FarmPaddockFarm H.Dry                     ***
## FarmPaddockFarm I.Dry                     ***
## FarmPaddockFarm A.Wet                     ***
## FarmPaddockFarm B.Wet                     ***
## FarmPaddockFarm D.Wet                     ***
## FarmPaddockFarm E.Wet                     ***
## FarmPaddockFarm F.Wet                     ***
## FarmPaddockFarm G.Wet                     ***
## FarmPaddockFarm H.Wet                     ***
## FarmPaddockFarm I.Wet                     ***
```

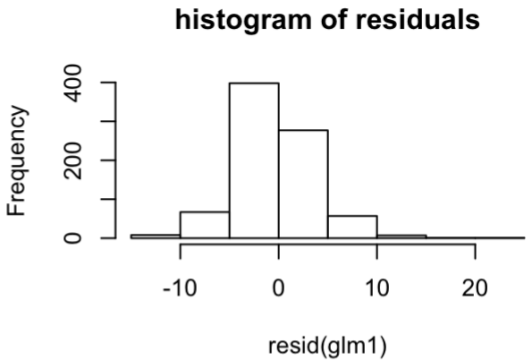
```
## HerbicideAminopyralid+triclopyr:MowNo Mow ***
## HerbicideFlumetsulam:MowNo Mow
## HerbicideMCPA:MowNo Mow                    .
## HerbicideMCPB:MowNo Mow                    ***
## HerbicideMCPB+bentazone:MowNo Mow          ***
## HerbicideNil:MowNo Mow                     ***
## HerbicideThifensulfuron-methyl:MowNo Mow   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 68695  on 815  degrees of freedom
## Residual deviance: 13124  on 784  degrees of freedom
## AIC: 17515
##
## Number of Fisher Scoring iterations: 6
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: buttercup$BareGrnd ~ Herbicide * Mow + (1 | FarmPaddock)
##    Data: buttercup
##
##      AIC      BIC   logLik deviance df.resid
##  17643.4  17723.4  -8804.7  17609.4      799
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -10.603  -2.497  -0.720   2.202  32.522
##
## Random effects:
##  Groups      Name        Variance Std.Dev.
##  FarmPaddock (Intercept) 2.021    1.422
## Number of obs: 816, groups:  FarmPaddock, 17
##
## Fixed effects:
```

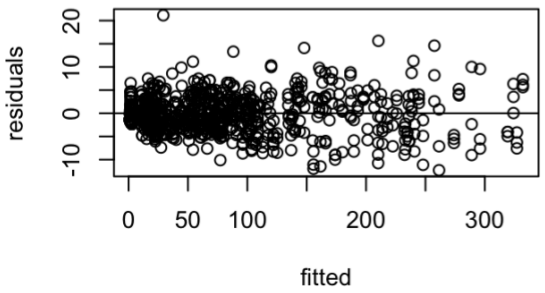
```
##                                          Estimate Std. Error z value Pr
(>|z|)
## (Intercept)                               4.13514    0.34511  11.982  <
 2e-16
## HerbicideAminopyralid+triclopyr           0.02406    0.01815   1.325
0.18505
## HerbicideFlumetsulam                     -0.28219    0.01970 -14.327  <
 2e-16
## HerbicideMCPA                            -0.15290    0.01900  -8.046 8.
56e-16
## HerbicideMCPB                            -0.36133    0.02015 -17.929  <
 2e-16
## HerbicideMCPB+bentazone                  -0.31731    0.01990 -15.948  <
 2e-16
## HerbicideNil                             -0.30095    0.01980 -15.198  <
 2e-16
## HerbicideThifensulfuron-methyl           -0.30728    0.01984 -15.489  <
 2e-16
## MowNo Mow                                -0.43329    0.02059 -21.042  <
 2e-16
## HerbicideAminopyralid+triclopyr:MowNo Mow  0.19439    0.02817   6.900 5.
19e-12
## HerbicideFlumetsulam:MowNo Mow            0.01949    0.03130   0.623
0.53341
## HerbicideMCPA:MowNo Mow                   0.05059    0.03005   1.683
0.09231
## HerbicideMCPB:MowNo Mow                   0.12472    0.03145   3.965 7.
33e-05
## HerbicideMCPB+bentazone:MowNo Mow         0.16294    0.03087   5.277 1.
31e-07
## HerbicideNil:MowNo Mow                   -0.24354    0.03306  -7.367 1.
74e-13
## HerbicideThifensulfuron-methyl:MowNo Mow  0.11809    0.03101   3.808
0.00014
##
## (Intercept)                              ***
## HerbicideAminopyralid+triclopyr
## HerbicideFlumetsulam                     ***
## HerbicideMCPA                            ***
## HerbicideMCPB                            ***
## HerbicideMCPB+bentazone                  ***
## HerbicideNil                             ***
## HerbicideThifensulfuron-methyl           ***
## MowNo Mow                                ***
```

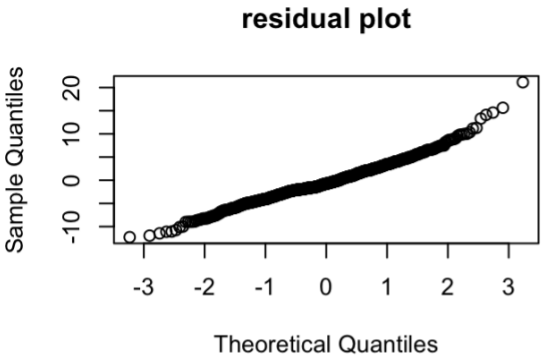
```
## HerbicideAminopyralid+triclopyr:MowNo Mow ***
## HerbicideFlumetsulam:MowNo Mow
## HerbicideMCPA:MowNo Mow                    .
## HerbicideMCPB:MowNo Mow                    ***
## HerbicideMCPB+bentazone:MowNo Mow          ***
## HerbicideNil:MowNo Mow                     ***
## HerbicideThifensulfuron-methyl:MowNo Mow   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation matrix not shown by default, as p = 16 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

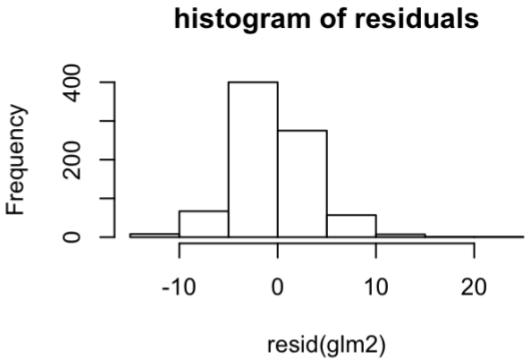
We compare the two models by model performances and AIC. Looking at the diagnostic plots, we cannot tell too much difference from the two models, and all the plots look fine, which also means they do not violate the model
asuumptions.

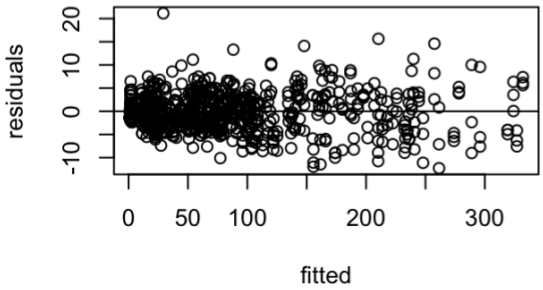












We then compare the models with anova() function.

```
## Data: buttercup
## Models:
## glm2: buttercup$BareGrnd ~ Herbicide * Mow + (1 | FarmPaddock)
## glm1: buttercup$BareGrnd ~ Herbicide * Mow + FarmPaddock
##      npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## glm2   17 17643 17723 -8804.7    17609
## glm1   32 17515 17665 -8725.3    17451 158.82 15  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see that the first model has a lower AIC and a lower deviance and the p-value for it is significant, which means it performs better than the other one. Comparing the outputs of the models, we can say all the predictor terms (Herbicide, Mow and FarmPaddock) we used to fit the models are useful for prediction and both of the two models perform fine as they do not violate any model assumptions. However, when we treat the Farm by Paddock interaction with fixed effects, the model performs slightly better.

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## Herbicide       7   42.4    6.06  17.354  < 2e-16 ***
## Mow             1   20.5   20.48  58.658 5.53e-14 ***
```

```
## FarmPaddock    16 1586.1   99.13 283.884  < 2e-16 ***
## Herbicide:Mow   7    4.5    0.64   1.824   0.0796 .
## Residuals     784  273.8    0.35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion of Question2

Similarly in question 2, we get the conclusion that Farm by Paddock effects is necessary by looking at a few plots of the data, then we fit a Generalised Linear model and a Generalised Linear model. Comparing the two models, we find out the generalised linear model performs slightly better.

QUESTION 3 FITTING A GENERALISED ADDITIVE MODEL

We start by loading the dataset, extract the Denmark data and change the date variable to an R date variable. We can check out the first six observations in the Denmark data.

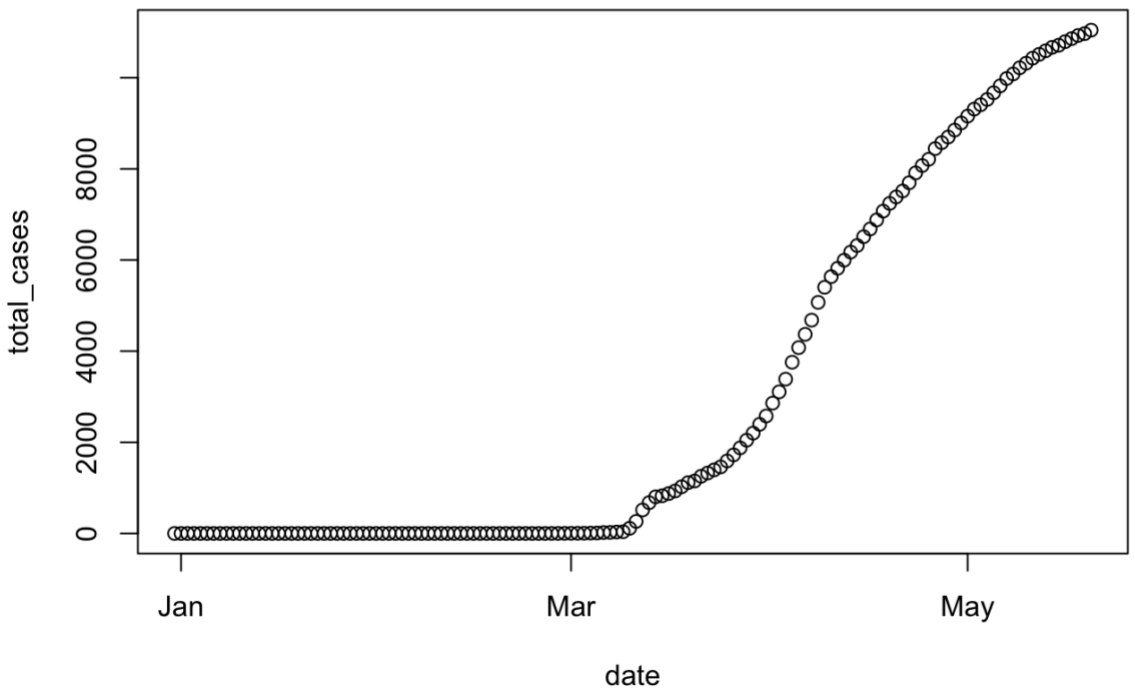
```
##      iso_code location       date total_cases new_cases total_deaths new
_deaths
## 4533      DNK  Denmark 2019-12-31           0         0            0
      0
## 4534      DNK  Denmark 2020-01-01           0         0            0
      0
## 4535      DNK  Denmark 2020-01-02           0         0            0
      0
## 4536      DNK  Denmark 2020-01-03           0         0            0
      0
## 4537      DNK  Denmark 2020-01-04           0         0            0
      0
## 4538      DNK  Denmark 2020-01-05           0         0            0
      0
##      total_cases_per_million new_cases_per_million total_deaths_per_mill
ion
## 4533                       0                     0
  0
## 4534                       0                     0
  0
## 4535                       0                     0
  0
## 4536                       0                     0
  0
## 4537                       0                     0
  0
```

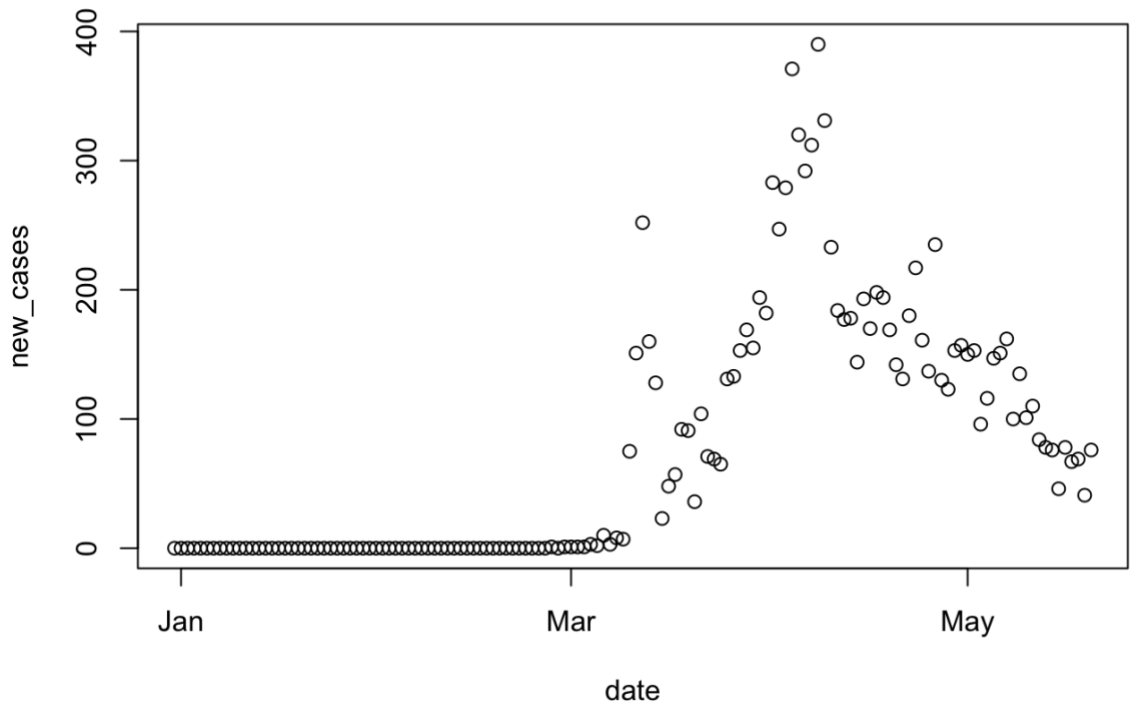
```
## 4538                       0                     0
  0
##      new_deaths_per_million total_tests new_tests new_tests_smoothed
## 4533                      0          NA        NA                 NA
## 4534                      0          NA        NA                 NA
## 4535                      0          NA        NA                 NA
## 4536                      0          NA        NA                 NA
## 4537                      0          NA        NA                 NA
## 4538                      0          NA        NA                 NA
##      total_tests_per_thousand new_tests_per_thousand
## 4533                       NA                     NA
## 4534                       NA                     NA
## 4535                       NA                     NA
## 4536                       NA                     NA
## 4537                       NA                     NA
## 4538                       NA                     NA
##      new_tests_smoothed_per_thousand tests_units stringency_index popula
tion
## 4533                              NA                          NA    579
2203
## 4534                              NA                           0    579
2203
## 4535                              NA                           0    579
2203
## 4536                              NA                           0    579
2203
## 4537                              NA                           0    579
2203
## 4538                              NA                           0    579
2203
##      population_density median_age aged_65_older aged_70_older gdp_per_c
apita
## 4533             136.52       42.3        19.677        12.325       466
82.51
## 4534             136.52       42.3        19.677        12.325       466
82.51
## 4535             136.52       42.3        19.677        12.325       466
82.51
## 4536             136.52       42.3        19.677        12.325       466
82.51
## 4537             136.52       42.3        19.677        12.325       466
82.51
```

```
## 4538             136.52       42.3        19.677        12.325       466
82.51

##      extreme_poverty cvd_death_rate diabetes_prevalence female_smokers
## 4533             0.2        114.767                6.41           19.3
## 4534             0.2        114.767                6.41           19.3
## 4535             0.2        114.767                6.41           19.3
## 4536             0.2        114.767                6.41           19.3
## 4537             0.2        114.767                6.41           19.3
## 4538             0.2        114.767                6.41           19.3
##      male_smokers handwashing_facilities hospital_beds_per_100k
## 4533         18.8                     NA                    2.5
## 4534         18.8                     NA                    2.5
## 4535         18.8                     NA                    2.5
## 4536         18.8                     NA                    2.5
## 4537         18.8                     NA                    2.5
## 4538         18.8                     NA                    2.5
```

We now draw graphs for each of the total_cases and new_cases vs date.



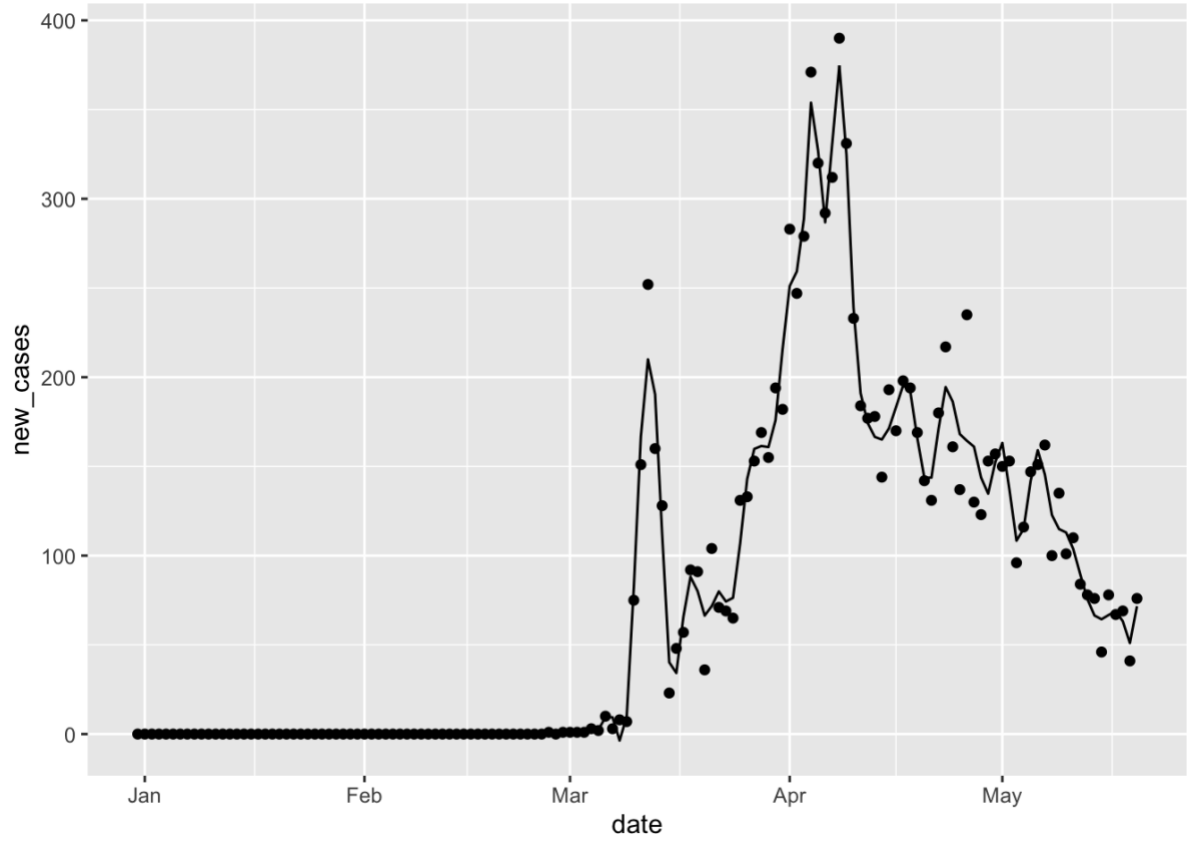


As can be seen in these two plots, both total_cases and new_cases remain zero till March, following an increase, but the new_cases drops in April. The relationships of cases and time are not linear, we might use a GAM to model the response.

For new_cases series we fit a generalized additive model (GAM) using the gam package with spar = 0.1, 0.3, 0.5, 0.7 and 0.9 values, we then plot the data and each of the fitted model.

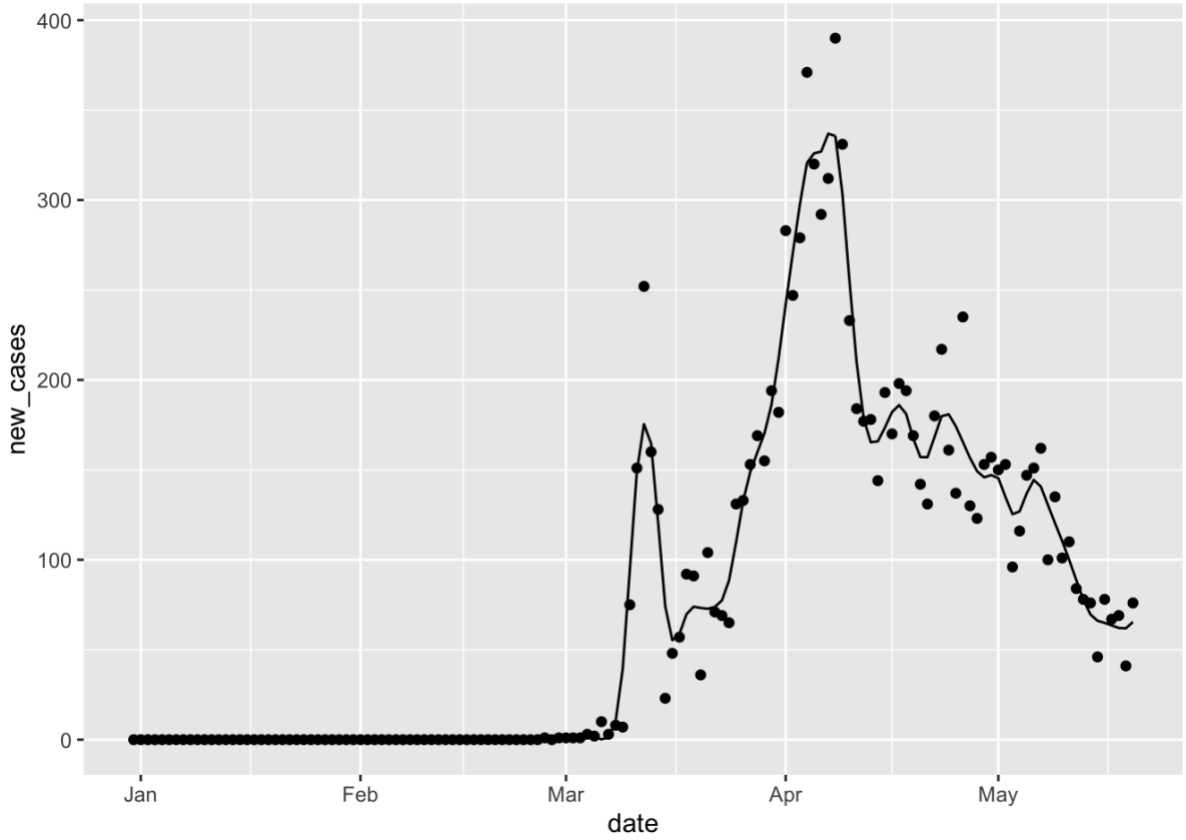
```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts a
rgument
## ignored
##
## Call: gam(formula = new_cases ~ s(date, spar = 0.1), data = covid_d)
## Deviance Residuals:
##        Min         1Q     Median         3Q        Max
## -3.398e+01 -2.629e+00  4.010e-13  1.853e+00  7.062e+01
##
## (Dispersion Parameter for gaussian family taken to be 344.0823)
##
##     Null Deviance: 1282875 on 141 degrees of freedom
## Residual Deviance: 24140.03 on 70.1577 degrees of freedom
## AIC: 1277.947
```

```
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## s(date, spar = 0.1)  1.000 558243  558243  1622.4 < 2.2e-16 ***
## Residuals           70.158  24140     344
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                      Npar Df Npar F     Pr(F)
## (Intercept)
## s(date, spar = 0.1)     69.8 29.149 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



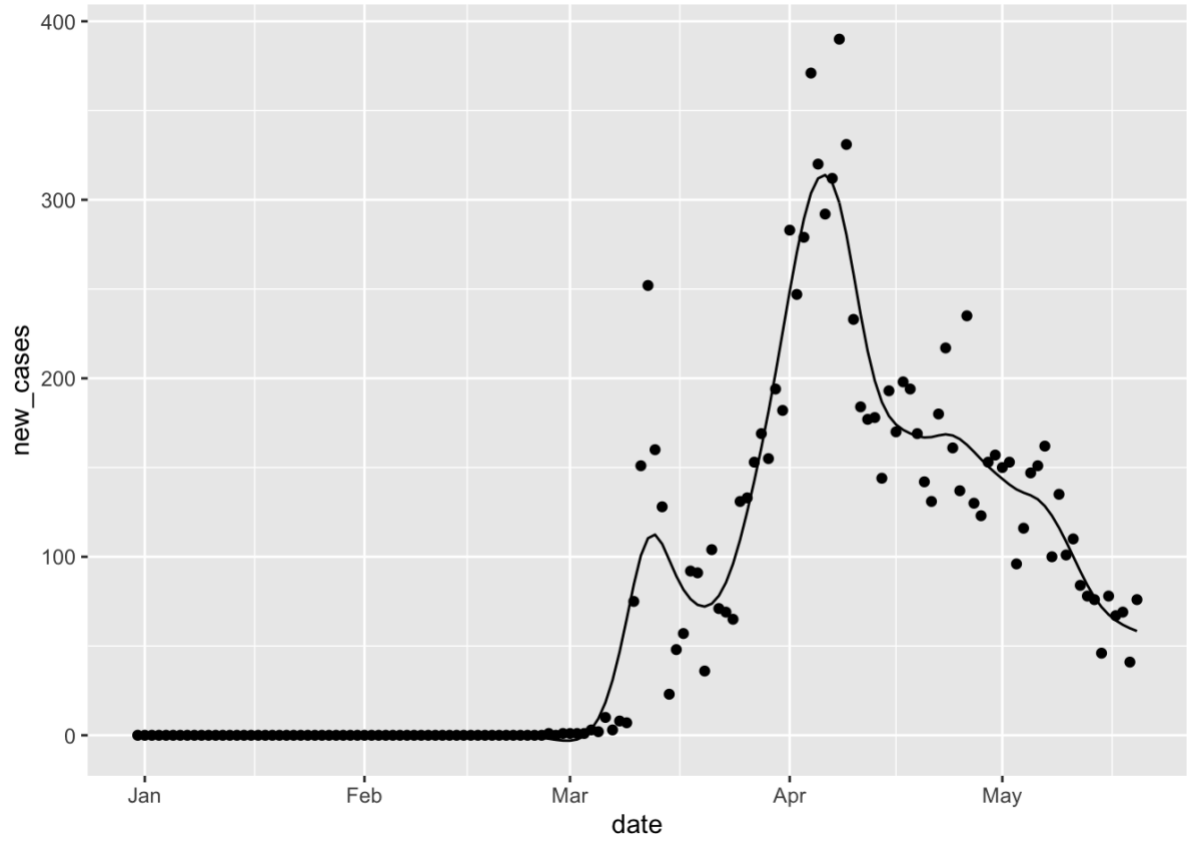

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts a
rgument
## ignored
```

```
##
## Call: gam(formula = new_cases ~ s(date, spar = 0.3), data = covid_d)
## Deviance Residuals:
##        Min         1Q     Median         3Q        Max
## -5.123e+01 -1.356e+00  1.101e-13  1.870e+00  7.659e+01
##
## (Dispersion Parameter for gaussian family taken to be 458.577)
##
##     Null Deviance: 1282875 on 141 degrees of freedom
## Residual Deviance: 44092.66 on 96.151 degrees of freedom
## AIC: 1311.504
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## s(date, spar = 0.3)  1.000 558243  558243  1217.3 < 2.2e-16 ***
## Residuals           96.151  44093     459
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                     Npar Df Npar F     Pr(F)
## (Intercept)
## s(date, spar = 0.3)    43.8 33.844 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



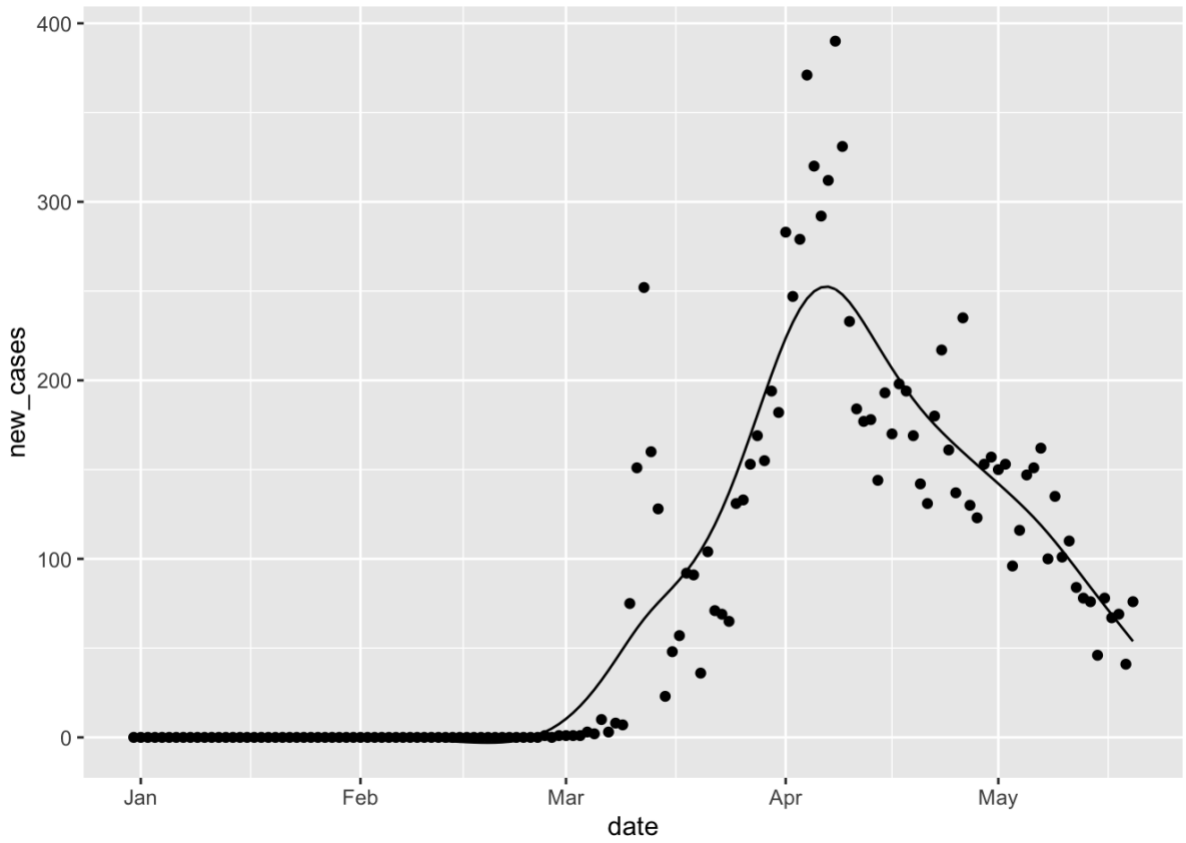

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts a
rgument
## ignored
##
## Call: gam(formula = new_cases ~ s(date, spar = 0.5), data = covid_d)
## Deviance Residuals:
##        Min         1Q     Median         3Q        Max
## -7.523e+01 -7.088e+00 -1.712e-08  3.248e+00  1.416e+02
##
## (Dispersion Parameter for gaussian family taken to be 766.8961)
##
##     Null Deviance: 1282875 on 141 degrees of freedom
## Residual Deviance: 92175.82 on 120.1934 degrees of freedom
## AIC: 1368.131
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                          Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## s(date, spar = 0.5)   1.00 558243  558243  727.93 < 2.2e-16 ***
## Residuals           120.19  92176     767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                     Npar Df Npar F     Pr(F)
## (Intercept)
## s(date, spar = 0.5)    19.8 41.637 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



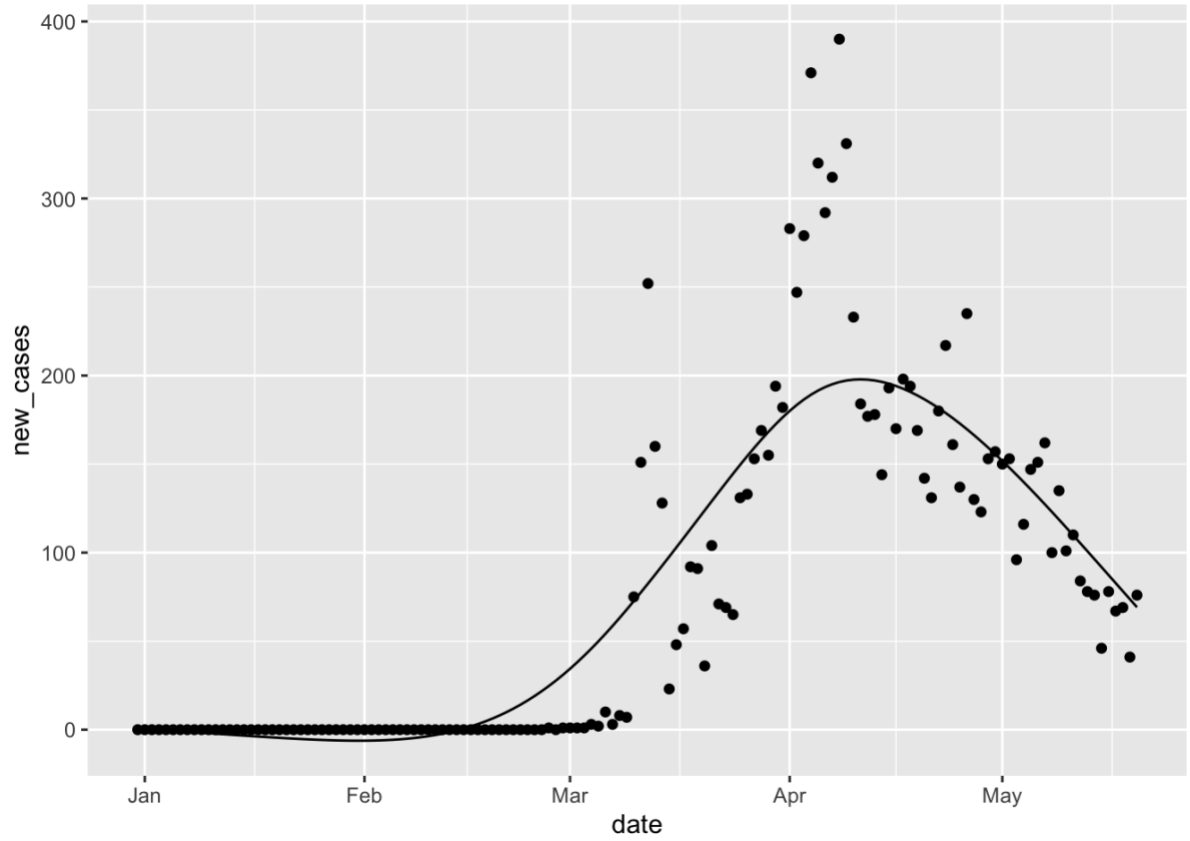

```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts a
rgument
## ignored
##
## Call: gam(formula = new_cases ~ s(date, spar = 0.7), data = covid_d)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -75.4079 -13.9738  -0.0454   2.4325 185.7272
```

```
##
## (Dispersion Parameter for gaussian family taken to be 1373.251)
##
##     Null Deviance: 1282875 on 141 degrees of freedom
## Residual Deviance: 181105.5 on 131.8808 degrees of freedom
## AIC: 1440.66
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                          Df Sum Sq Mean Sq F value    Pr(>F)
## s(date, spar = 0.7)    1.00 558243  558243  406.51 < 2.2e-16 ***
## Residuals            131.88 181105    1373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                      Npar Df Npar F     Pr(F)
## (Intercept)
## s(date, spar = 0.7)      8.1 48.748 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```




```
## Warning in model.matrix.default(mt, mf, contrasts): non-list contrasts a
rgument
## ignored
##
## Call: gam(formula = new_cases ~ s(date, spar = 0.9), data = covid_d)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -87.8239 -21.4328   0.2921   5.4420 193.6664
##
## (Dispersion Parameter for gaussian family taken to be 2088.26)
##
##     Null Deviance: 1282875 on 141 degrees of freedom
## Residual Deviance: 286149.6 on 137.0277 degrees of freedom
## AIC: 1495.322
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##                          Df Sum Sq Mean Sq F value    Pr(>F)
```

```
## s(date, spar = 0.9)   1.00 558243  558243  267.32 < 2.2e-16 ***
## Residuals           137.03 286150    2088
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##                     Npar Df Npar F     Pr(F)
## (Intercept)
## s(date, spar = 0.9)       3 70.645 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Having all the models fitted and the fitted data plotted, we can see as the spar values(smoothing parameters) increase, models tend to better smooth the data and demonstrate general trends. For the very first model, it almost joints dots together and does not smooth or show any trend, therefore it does not seem to be a good choice. When the spar value is 0.3, the model captures the general trend

Discussion of Question3

In this question, we investigate the covid cases for Denmark dataset. We explore the data by plotting total_cases and new_cases against time, this gives us a better idea of what kind of model can be

approximate. We then fit the data with generalized additive modoels using different smoothing parameters, compare the summary outputs and plots to choose the one that performs ‘best’ in this case.