Assignment – 2

STAT30250 - Advanced Predictive Analytics

Meghashree Madhava Rao Ramachandrahosur (**21200301**)

Question 1

List of references of Generalised Additive Model applied to COVID-19 data:

1. Liu, K. T., Gong, Y. N., Huang, C. G., Huang, P. N., Yu, K. Y., Lee, H. C., ... & Shih, S. R. (2022). Quantifying Neutralizing Antibodies in Patients with COVID-19 by a Two-Variable Generalized Additive Model. *Msphere*, *7*(1), e00883-21.

2. Dikaya, L. A., Avanesian, G., Dikiy, I. S., Kirik, V. A., & Egorova, V. A. (2021). How personality traits are related to the attitudes toward forced remote learning during COVID-19: predictive analysis using generalized additive modeling. *Front. Educ*, *6*(629213), 10-3389.

3. Fotiadis, A., Polyzos, S., & Huan, T. C. T. (2021). The good, the bad and the ugly on COVID-19 tourism recovery. *Annals of Tourism Research*, *87*, 103117.

4. Gupta, A., Banerjee, S., & Das, S. (2020). Significance of geographical factors to the COVID-19 outbreak in India. *Modeling earth systems and environment*, *6*(4), 2645-2653.

5. Izadi, F. (2022). Generalized additive models to capture the death rates in Canada COVID-19. In *Mathematics of Public Health* (pp. 153-171). Springer, Cham.

Load Required Libraries

```
library("psych")
library(ggplot2)
library(mgcv)
library(car)
library(lmtest)


# set the working directory and load the data set
setwd("C:/Users/DELL/OneDrive/Documents/SEM-2/PA/APA-3")
irl <- read.csv("IRL.csv")
```

```
data <- irl
```

Data Pre-processing

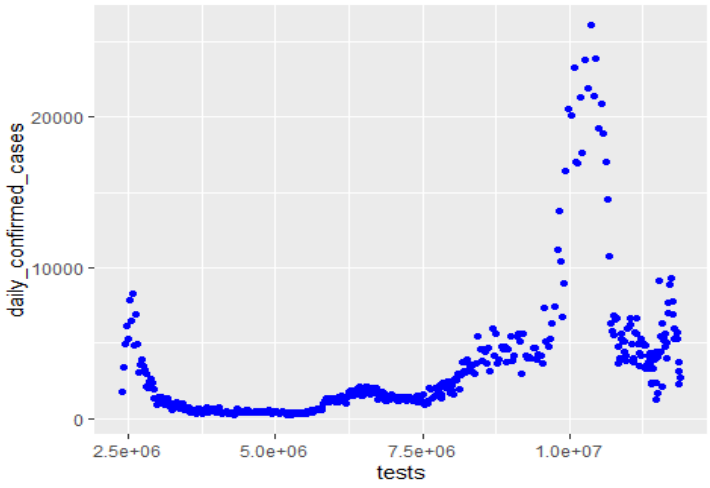
```
# format the variables into the right form for analysis
data$date <- as.Date(data$date, format = "%d/%m/%Y")

# convert categorical variables into factors
data$school_closing <- as.factor(data$school_closing)
data$transport_closing <- as.factor(data$transport_closing)
data$gatherings_restrictions <- as.factor(data$gatherings_restrictions)

str(data)

## 'data.frame':    460 obs. of  10 variables:
##  $ date                   : Date, format: "2021-01-01" "2021-01-02" ...
##  $ daily_confirmed_cases  : int  1754 3394 4962 6110 5325 7836 6521 8248 4
842 6888 ...
##  $ tests                  : int  2402399 2423245 2451788 2472359 2492267 2
520637 2549247 2576566 2606538 2631023 ...
##  $ vaccines               : int  3193 4208 4231 5420 8953 15732 27458 4013
8 48216 52997 ...
##  $ people_fully_vaccinated: int  22 22 25 28 37 45 53 58 64 68 ...
##  $ hosp                   : int  508 581 673 744 817 921 1022 1153 1285 14
26 ...
##  $ icu                    : int  50 56 65 73 76 89 101 107 121 128 ...
##  $ school_closing         : Factor w/ 4 levels "0","1","2","3": 4 4 4 4 4
4 4 4 4 4 ...
##  $ gatherings_restrictions: Factor w/ 5 levels "0","1","2","3",..: 5 5 5 5
5 5 5 5 5 5 ...
##  $ transport_closing      : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2
2 ...

# Find for any NA values
apply(is.na(data), 2, which)

## integer(0)
```

There was no missing data or ‘NA’ in the data set given. All the data is now in useful form.

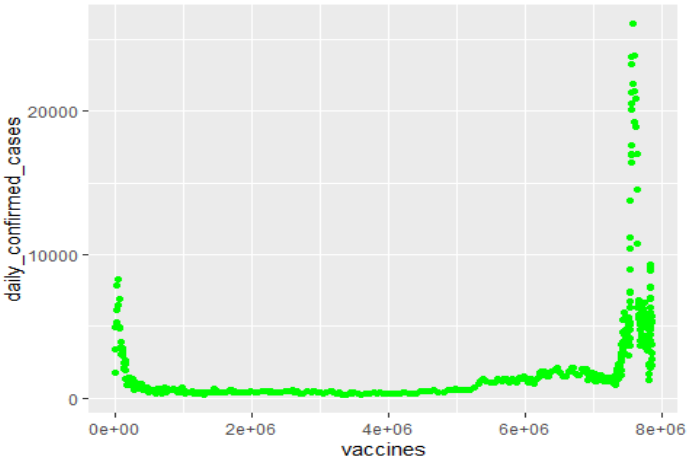
Question 2

Plot of Continuous Variables with Target

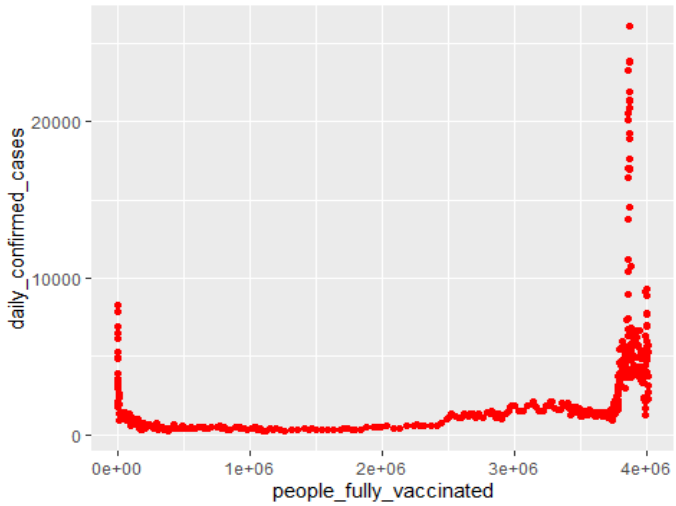
```
#Plot the number of confirmed daily cases of COVID-19 in Ireland

par(mfrow = c(3,2))

ggplot(aes(x=tests,y=daily_confirmed_cases),data=data) + geom_point(color = "
blue")
```



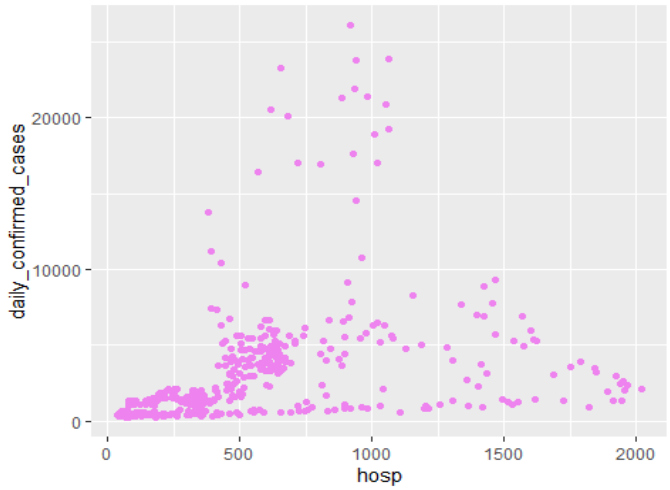

```
ggplot(aes(x=vaccines,y=daily_confirmed_cases),data=data) + geom_point(color
= "green")
```



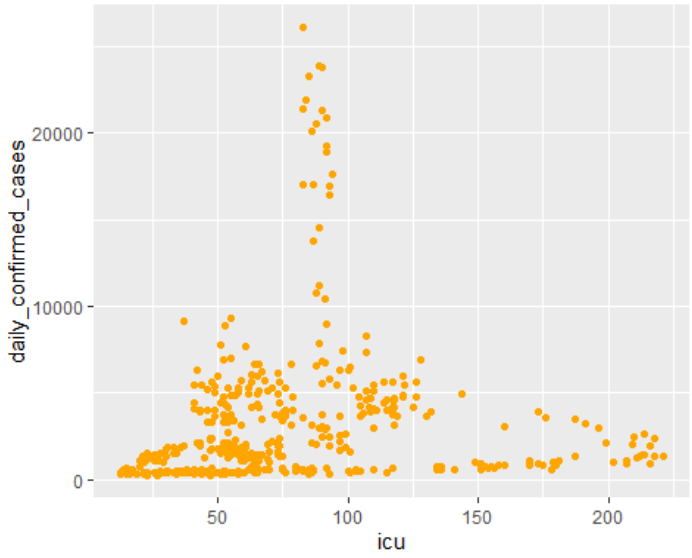

```
ggplot(aes(x=people_fully_vaccinated,y=daily_confirmed_cases),data=data) + ge
om_point(color = "red")
```



```
ggplot(aes(x=hosp,y=daily_confirmed_cases),data=data) + geom_point(color = "violet")
```




```
ggplot(aes(x=icu,y=daily_confirmed_cases),data=data) + geom_point(color = "orange")
```



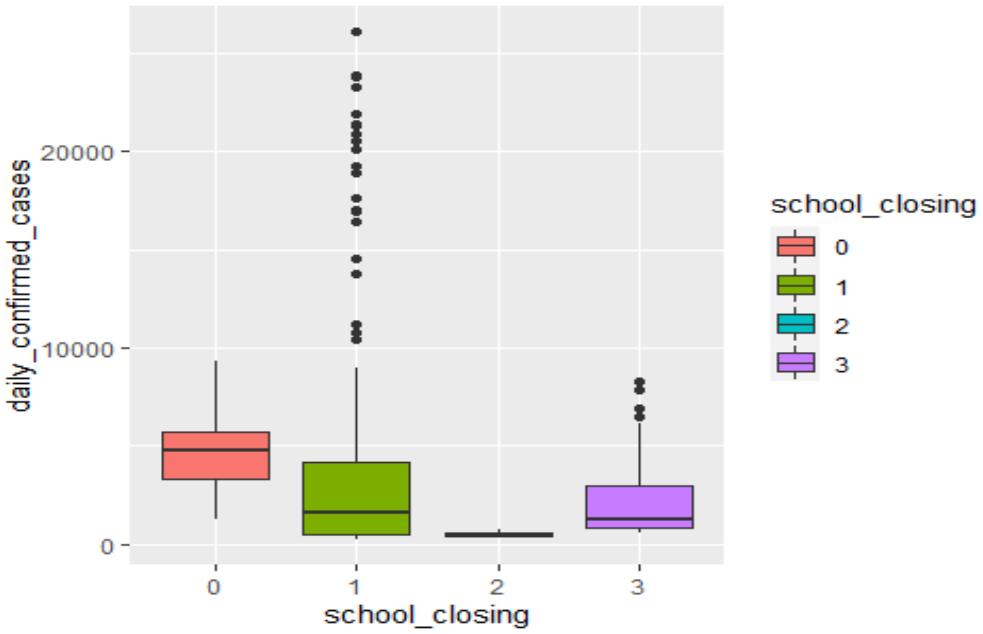
All the continuous predictor variables are not linearly related to the response variable. The first 3 plots obtained from plotting variables tests, vaccines and number of people fully vaccinated, we can say vaguely that they resemble log-normal distribution with μ- mean and σ – standard deviation different from 0.

The variables *hosp* and *icu* are very vaguely Poisson distributed with respect to the response variable. This indicates that there are other factors accounting to the daily confirmed number of Covid-19 cases. We will have to use smooth to effectively account for these variables by reducing the penalty.

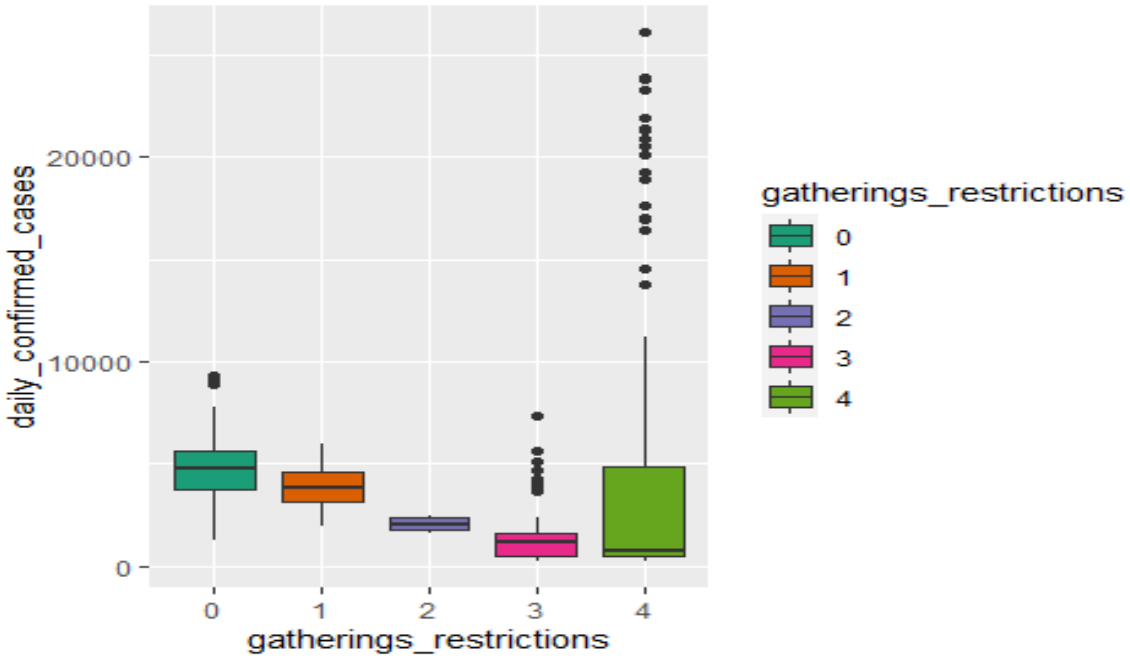
Question 3

Plot of Categorical Variables with Target

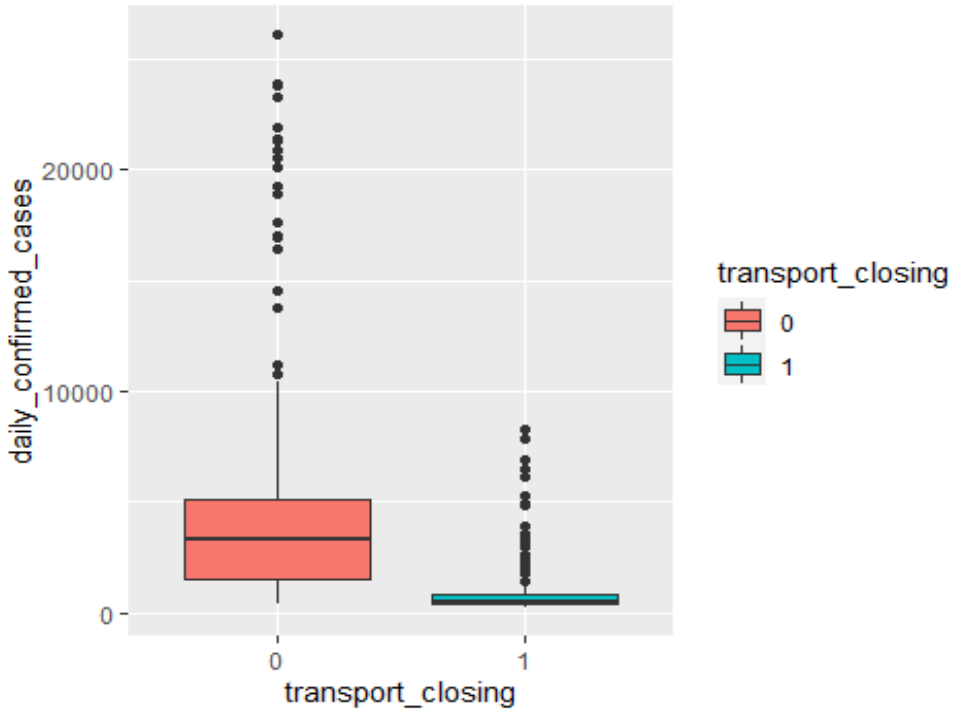
```
ggplot(data, aes(x=school_closing, y=daily_confirmed_cases, fill = school_clo
sing)) + geom_boxplot()
```




```
ggplot(data, aes(x=gatherings_restrictions, y=daily_confirmed_cases, fill = g
atherings_restrictions)) + geom_boxplot()+ scale_fill_brewer(palette="Dark2")
```



```
ggplot(data, aes(x=transport_closing, y=daily_confirmed_cases, fill = transpo
rt_closing)) + geom_boxplot()
```



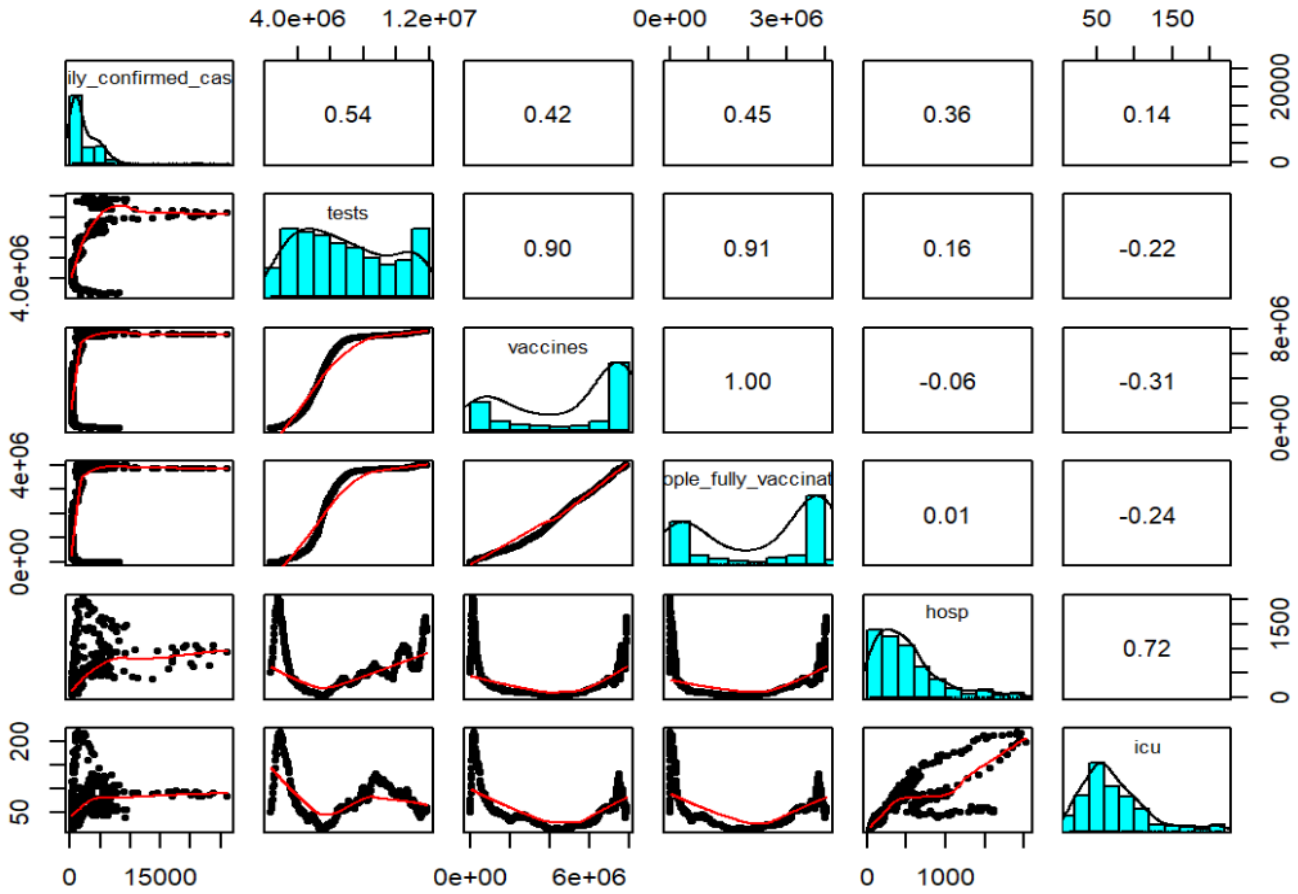
We use boxplot to observe the distributions of the categorical variables that describe the measures taken by government where 0 indicating no measures, highest number – indicating complete strict regulations and in between range indicating measures partially rolled out in stages.

As the requirements become more stringent, all restrictions strongly suggest a decrease in the average number of daily Covid-19 cases. The restriction of gathering 10 people or less has a highly positively skewed distribution with the response variable, as do the other variables where schools were recommended closing or all schools were recommended opening with alterations resulting in significant differences compared to non-Covid-19 operations and where travel was not restricted.

The distribution of covid-19 cases when transport regulations were applied has a sharp bell curve with many outliers.

All the Categorical variables seem to have good skewed relationship with the response variable implying no smooth / penalty needed for analysis.

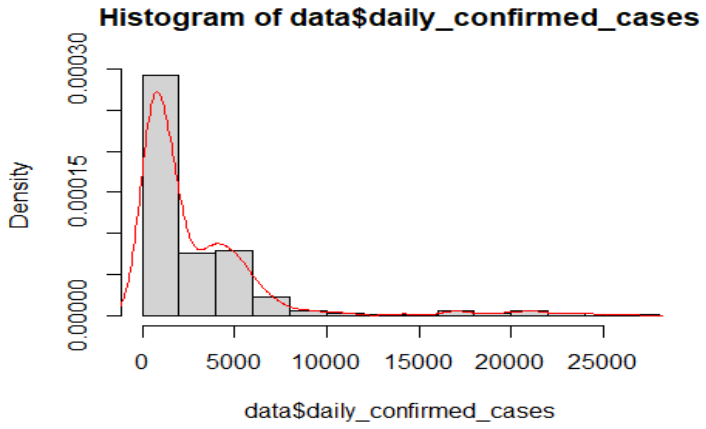
```
# The pairs panel plot provides the distribution of continuous variables with
the response variable on left triangular matrix and correlation between varia
bles on the upper triangular plot

pairs.panels(data[,2:7], smooth = TRUE, scale = FALSE, density=TRUE, ellipses
=FALSE)
```



The pairs plot acts as a validation method to understand that the continuous variables are not linearly distributed, and if it were to be modelled as is, the correlation between the variables is not great to obtain a good linear predictive model.

Response variable distribution

```
hist(data$daily_confirmed_cases, freq = FALSE)
lines(x = density(data$daily_confirmed_cases), col = "red")
```



By looking at the distribution of response variable, we can say that it is Poisson distributed. Population count data is discrete and over dispersed in this example dataset. Overdispersion is the presence of greater variability in a data set than would be expected based on a given statistical model. This can be found by fitting a gam() model without any method and analyzing its residuals. This is done in later stages in Q4.

Question 4

We shall consider the day of first observation to be day 0, next day 1 and so on. Let us add a Day column to the data set for inclusion of date as a variable. Next, let us scale the data for variables concerned with population like, *tests,vaccines* and *people_fully_vaccinated*. We will convert the population to per million quantity for better correlation and interpretation purposes.

```
data$Day <- seq(0,459)

# Summary of continuous data
summary(data[,2:7])
```

```
##  daily_confirmed_cases     tests              vaccines
##  Min.   :  242.0       Min.   : 2402399   Min.   :   3193
##  1st Qu.:  572.8       1st Qu.: 4413757   1st Qu.:1521535
##  Median : 1493.5       Median : 6509580   Median :6648629
##  Mean   : 3045.5       Mean   : 6910425   Mean   :4907704
##  3rd Qu.: 4047.8       3rd Qu.: 9353711   3rd Qu.:7512390
##  Max.   :26122.0       Max.   :11875799   Max.   :7846099

##  people_fully_vaccinated      hosp             icu
##  Min.   :     22         Min.   :  38.0   Min.   : 13.00
##  1st Qu.: 439813         1st Qu.: 196.8   1st Qu.: 44.00
##  Median :3255558         Median : 424.5   Median : 63.00
##  Mean   :2372607         Mean   : 536.0   Mean   : 74.92
##  3rd Qu.:3846378         3rd Qu.: 669.2   3rd Qu.: 93.00
##  Max.   :4013615         Max.   :2020.0   Max.   :221.00
```

```
data$people_fully_vaccinated <- as.numeric(data$people_fully_vaccinated)/10^6
data$tests <- as.numeric(data$tests)/10^6
data$vaccines <- as.numeric(data$vaccines)/10^6
```

We can fit the model with *link="log"* either with Negative Binomial or Gamma distribution of the response variable.

Negative binomial regression is for modeling count variables, usually for over-dispersed count outcome variables. Since, response variable is count/ discrete data/ whole numbers, we use Poisson distribution to model the data. But the fit model is over dispersed. We check for dispersion by either plotting the QQ-plot or by dividing the sum of residuals by the degrees of freedom of residuals. We obtain a value of 141.9 which very high compared to one.

```
# Fit the plain model
```

```
mod_pois <- gam(daily_confirmed_cases ~ s(Day)+s(tests)+s(vaccines)+s(people_
fully_vaccinated)+s(hosp)+s(icu), data = data, family = poisson(link = "log")
, method = "REML")

# check for over dispersion
sum(residuals(mod_pois, type = "pearson")^2) / df.residual(mod_pois)

## [1] 141.9576
```

Since the value of dispersion with Poisson model is high, we fit the model with Negative Binomial distribution to account for the over dispersion. The variance and mean of the distribution does not have to be equal for Negative Binomial distribution.

We can model the data with Gamma distribution too. But Gamma distribution is used to model continuous data. Continuous data is not whole numbers. It has freedom to account for any number in between the range.

When both models are plotted out, both provide similar results with effective degrees of freedom. *'REML'* does not indicate much difference between the two models. Keeping in mind that we also hope to get the prediction in whole/discrete numbers, we go ahead and work with Negative binomial family with *'log'* link function.

By modelling with negative binomial distribution, we also overcome over dispersion (residual/df ~ 1). Whereas, the gamma model has possible under dispersion (residual/df ~ 0.05).

```
# Fit the gam model with Negative Binomial
mod_nb <- gam(daily_confirmed_cases ~ s(Day)+s(tests)+s(vaccines)+s(people_fu
lly_vaccinated)+s(hosp)+s(icu), data = data, family = nb(link = "log"), metho
d = "REML")

# Fit the gam model with Gamma distribution
mod_gam <- gam(daily_confirmed_cases ~ s(Day)+s(tests)+s(vaccines)+s(people_f
ully_vaccinated)+s(hosp)+s(icu), data = data, family = Gamma(link = "log"), m
ethod = "REML")

summary.gam(mod_nb)
##
## Family: Negative Binomial(21.934)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day) + s(tests) + s(vaccines) + s(people_fully_v
accinated) +
##     s(hosp) + s(icu)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.41618    0.01006   736.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Approximate significance of smooth terms:
##                               edf Ref.df Chi.sq  p-value
## s(Day)                      6.723  7.425 16.212   0.0332 *
## s(tests)                    3.569  4.337  7.882   0.1151
## s(vaccines)                 8.415  8.672 81.109  < 2e-16 ***
## s(people_fully_vaccinated) 5.982  7.069 50.532  < 2e-16 ***
## s(hosp)                     2.163  2.739 37.522 1.45e-06 ***
## s(icu)                      6.750  7.863 60.900  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.92   Deviance explained = 96.6%
## -REML = 3427.2  Scale est. = 1         n = 460
```

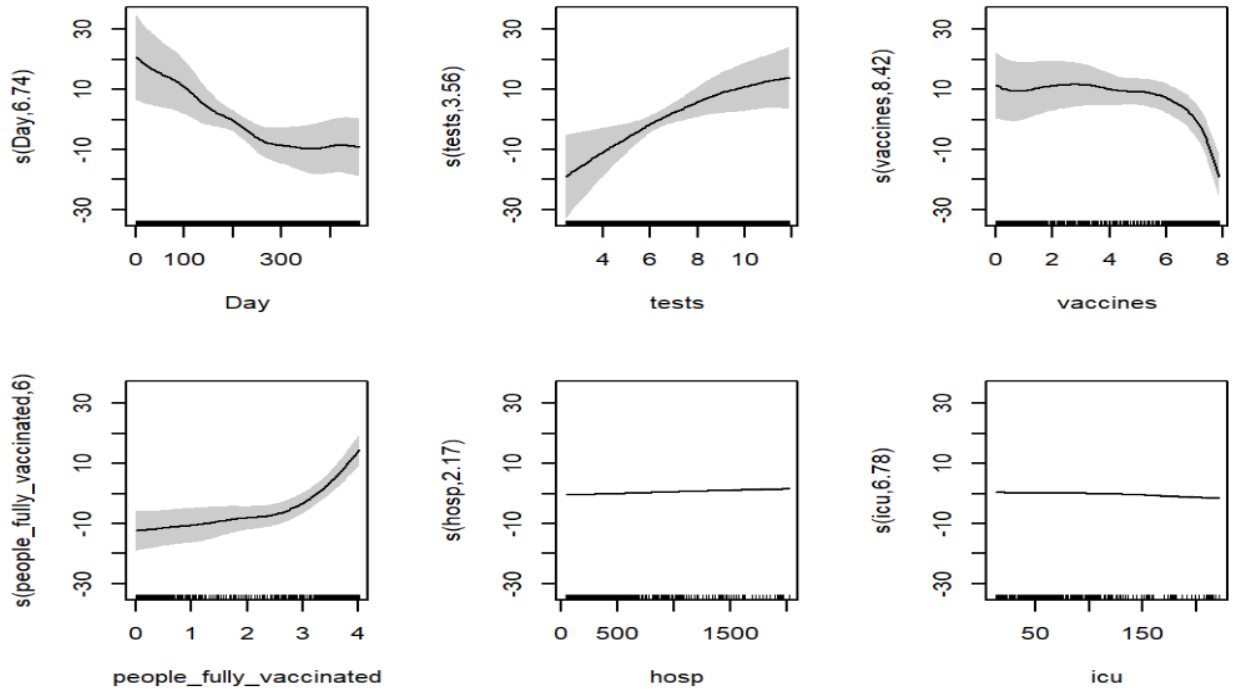
```
summary.gam(mod_gam)
```

```
##
## Family: Gamma
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day) + s(tests) + s(vaccines) + s(people_fully_v
accinated) + s(hosp) + s(icu)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.416166   0.009797     757   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                               edf Ref.df      F  p-value
## s(Day)                      6.738  7.441  2.318   0.0253 *
## s(tests)                    3.562  4.331  1.919   0.0994 .
## s(vaccines)                 8.420  8.676  9.757  < 2e-16 ***
## s(people_fully_vaccinated) 6.004  7.090  7.410  < 2e-16 ***
## s(hosp)                     2.170  2.748 14.106 3.95e-07 ***
## s(icu)                      6.784  7.892  8.118  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =   0.92   Deviance explained = 96.6%
## -REML = 3425.9  Scale est. = 0.044152  n = 460
```

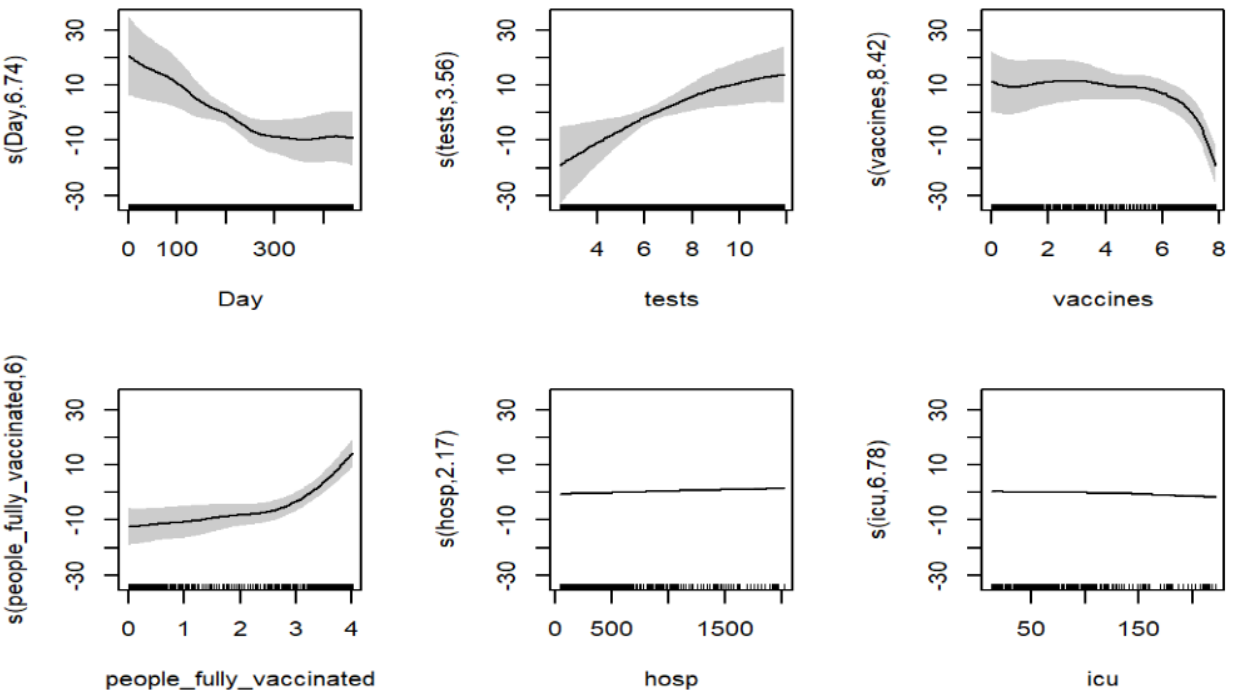
```
# Obtain AIC values for model comparison
AIC(mod_nb,mod_gam)
```

```
##               df      AIC
## mod_nb  40.10619 6748.424
## mod_gam 40.17901 6745.476
```

```
plot.gam(mod_nb, shade = TRUE, pages = 1)
```




```
plot.gam(mod_gam, shade = TRUE, pages = 1)
```




```
# check for over dispersion - value close to 1
sum(residuals(mod_nb, type = "pearson")^2) / df.residual(mod_nb)

## [1] 0.9534601
```

```
sum(residuals(mod_gam, type = "pearson")^2) / df.residual(mod_gam)

## [1] 0.0441516 # under dispersion
```

Question 5

$$\log\{E(daily\_confirmed\_cases\ _i)\} = 7.4162 + f_1(Day_i) + f_2(tests_i) + f_3(vaccines_i) + f_4(people\_fully\_vaccinated_i) + f_5(hosp_i) + f_6(icu_i),$$

$$\boldsymbol{daily\_confirmed\_cases\ _i \sim Negative\_Binomial}()$$

Where, f_i is the smooth functions applied to the variables and the median value of daily_confirmed_cases follow a *'Negative binomial'* distribution. Each term is incorporated into the model as a smooth function.

Question 6

We have 6 variable smooths and we want to estimate the model coefficients by minimizing,

$$||y - B_1c_1 - B_2c_2 - B_3c_3 - B_4c_4 - B_5c_5 - B_6c_6||^2 + \lambda_1c_1'R_1c_1 + \lambda_2c_2'R_2c_2 + \lambda_3c_3'R_3c_3 + \lambda_4c_4'R_4c_4 + \lambda_5c_5'R_5c_5 + \lambda_6c_6'R_6c_6$$

Where,
y is the response variable vector
$\lambda_1, \lambda_2 \dots \dots \lambda_6$ are the smoothing parameters which control the weight to be given to the objective of making $f_1 \dots \dots f_6$ smooth, relative to the objective of closely fitting the response data.
$B_1 \dots . . B_6$are basis functions.

We can simplify the equation,

$Let, B = [B_1 \dots \dots B_6]$

$Let\ R\ be\ the\ (K_1 + \cdots + K_6) \times (K_1 + \cdots + K_6) matrix$

$$R = \begin{matrix} \lambda_1R_1 & \dots & 0 \\ : & \dots & : \\ 0 & \dots & \lambda_6R_6 \end{matrix}$$

$$||y - Bc^*||^2 + c^*Rc^*$$

$$\hat{c^*} = (B'B + R)^{-1}By$$

Where,

$\hat{c^*} = [c_1, c_2, c_3, c_4, c_5, c_6]$ is the vector estimate of all coefficients,

$R_1 \dots R_6$ are the roughness penalty matrices with and K is the knots

Question 7

Tuning k for each smooth term

We determine approximate k values does not restrict the model basis dimension too much. We observe the edf and Ref. df values from the summary of the model. If the edf value is very close to the Ref. df, generally we can start with doubling the k value to observe any differences and check the model with gam.check(). Default k-value is set to 10.

Predictor variables *Day* and *vaccine* have very close values. Hence we start with doubling k-values to check it's influence on the model output. Once we bring up the k-value of *Day* variable to 30, the model passes all the checks and there are no more low k-index values for the predictors.

```
gam.check(mod_nb)
```

```
##
## Method: REML   Optimizer: outer newton
## full convergence after 8 iterations.
## Gradient range [-8.640166e-05,0.0002037842]
## (score 3427.181 & scale 1).
## Hessian positive definite, eigenvalue range [0.1809577,203.1779].
## Model rank =  55 / 55
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                              k'  edf k-index p-value
## s(Day)                     9.00 6.72    0.74  <2e-16 ***
## s(tests)                   9.00 3.57    0.74  <2e-16 ***
## s(vaccines)                9.00 8.42    0.74  <2e-16 ***
## s(people_fully_vaccinated) 9.00 5.98    0.74  <2e-16 ***
## s(hosp)                    9.00 2.16    0.98    0.39
## s(icu)                     9.00 6.75    0.97    0.20
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Fit the gam model with 'k' value
modk <- bam(daily_confirmed_cases ~ s(Day, k=30)+s(tests)+s(vaccines)+s(peopl
e_fully_vaccinated)+s(hosp)+s(icu), data = data, family = nb(link = "log"))

gam.check(modk)
```

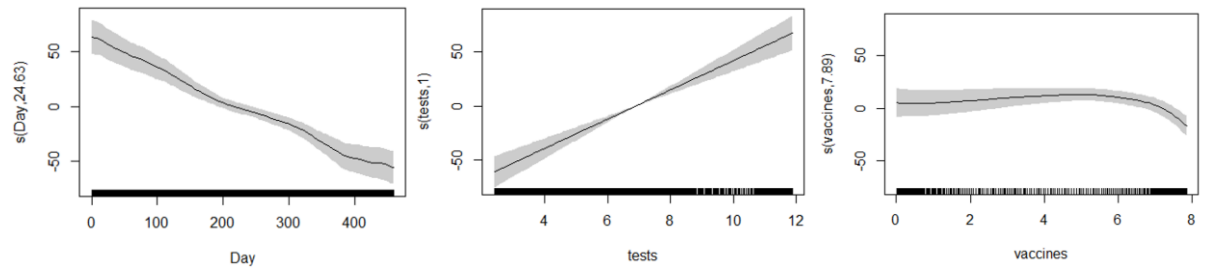
```
##
## Method: fREML   Optimizer: perf newton
## full convergence after 33 iterations.
## Gradient range [-8.222482e-07,2.85536e-07]
## (score 784.9015 & scale 1).
## Hessian positive definite, eigenvalue range [4.023795e-07,10.01114].
## Model rank =  75 / 75
##
```

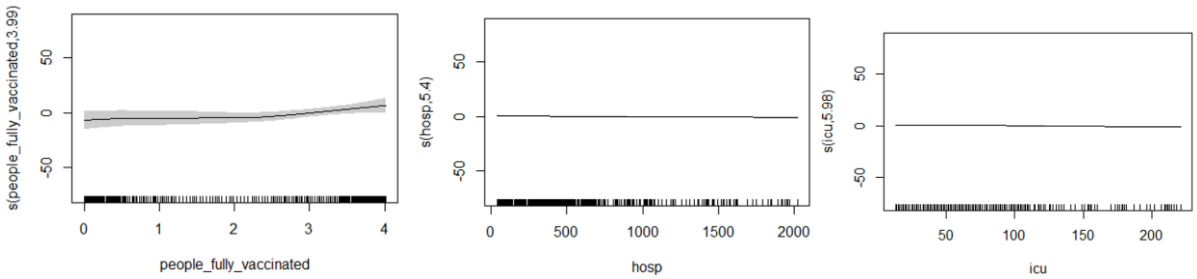
```
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                               k'   edf k-index p-value
## s(Day)                     29.00 24.63    1.00    0.42
## s(tests)                    9.00  1.00    1.00    0.48
## s(vaccines)                 9.00  7.89    1.00    0.44
## s(people_fully_vaccinated)  9.00  3.99    1.00    0.44
## s(hosp)                     9.00  5.40    1.02    0.64
## s(icu)                      9.00  5.98    0.99    0.35
```

```
summary,gam(modk)
```

```
## Family: Gamma
## Link function: log

## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + s(tests) + s(vaccines) +
##    s(people_fully_vaccinated) + s(hosp) + s(icu)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
##  (Intercept) 7.409100   0.008226   900.7   <2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Approximate significance of smooth terms:
##                               edf Ref.df      F  p-value
## s(Day)                     24.496 26.338 12.259  < 2e-16 ***
## s(tests)                    1.000  1.000 63.433  < 2e-16 ***
## s(vaccines)                 7.804  8.254  5.523 1.35e-06 ***
## s(people_fully_vaccinated)  3.755  4.742  1.224  0.27786
## s(hosp)                     5.129  6.291  2.554  0.04188 *
## s(icu)                      5.620  6.904  3.286  0.00212 **
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## R-sq.(adj) =  0.955   Deviance explained = 97.7%
## -REML = -33.51  Scale est. = 0.031125  n = 460
```

```
plot.gam(modk, shade = T)
```





Question 8

Wald tests in the summary table is used to find approximate significance of the smooth terms:
Hypothesis tests:

$$H_0: f(X_p + 1) = 0$$
$$H_a: f(X_p + 1) \neq 0$$

If the p-value of the test is <0.05 reject the null hypothesis. We conclude that there is evidence that $f(X_p + 1) \neq 0$ and thus there is a non-linear relationship between $X_p + 1$ and $E(Y_i)$.

Looking back at the summary of the previous fit model, the predictors, '*Day'*, '*tests'*, '*vaccines'*, '*hosp'* and '*icu'* are significant smooth terms at 95% confidence level since their p-values are all <0.05. Predictor *'people_fully_vaccinated'* is not a significant smooth term as its p-value 0.277 >0.05.

Question 9

According to the Wald tests in the summary table, F-value of '*tests'* predictor is the highest, hence indicating it as the most influential on the number of confirmed daily cases of COVID-19 in Ireland from 01/01/2021 to 05/04/2022. But we can also observe that the effective degrees of freedom for s(test) indicates that the term does not require smooth and can be deemed as linear term.
Hence, the next most significant variable, '*Day'* is considered the most influential on the number of confirmed daily cases of COVID-19 in Ireland from 01/01/2021 to 05/04/2022.

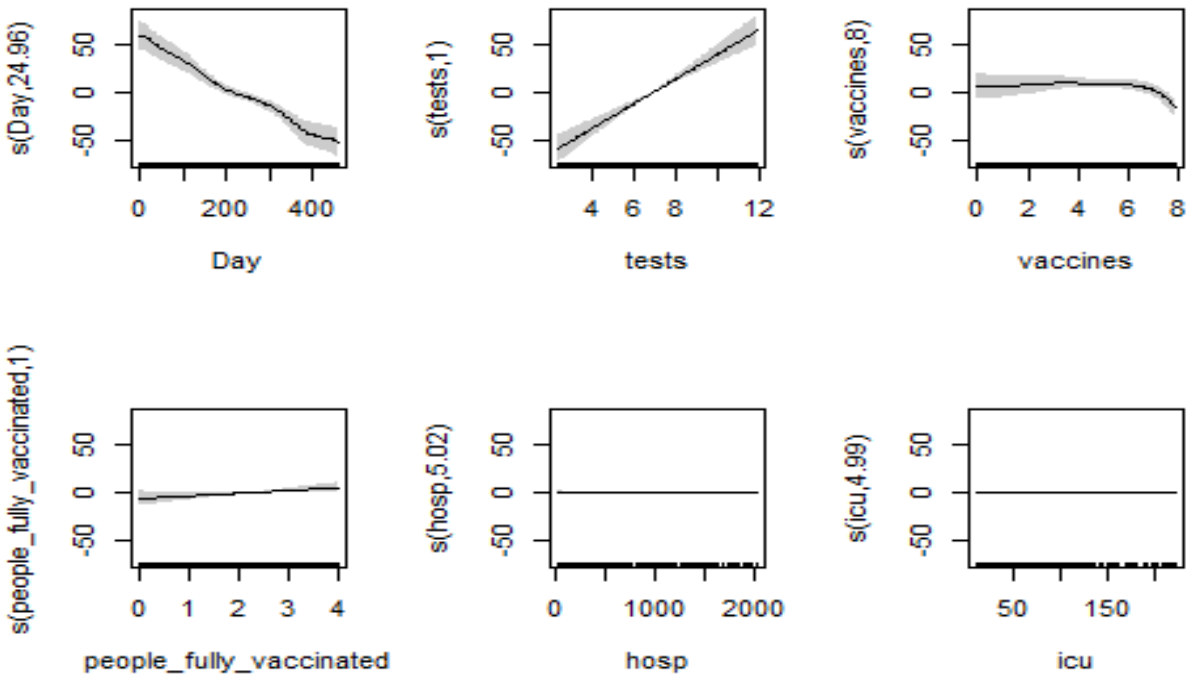
Question 10

Refit model with method = "ML"

```
# Fit the gam model
mod_refit <- bam(daily_confirmed_cases ~ s(Day, k=30)+s(tests)+s(vaccines)+s(
people_fully_vaccinated)+s(hosp)+s(icu), data = data, family = nb(link = "log
"), method = "ML")

summary.gam(mod_refit)

##
## Family: Negative Binomial(34.636)
## Link function: log
```

```
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + s(tests) + s(vaccines) +
##     s(people_fully_vaccinated) + s(hosp) + s(icu)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.40955    0.00806   919.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                               edf Ref.df      F p-value
## s(Day)                     24.961 26.785 24.287 < 2e-16 ***
## s(tests)                    1.000  1.000 65.803 < 2e-16 ***
## s(vaccines)                 7.997  8.460  5.746   1e-06 ***
## s(people_fully_vaccinated)  1.000  1.000  3.191 0.07476 .
## s(hosp)                     5.017  6.182  2.664 0.02812 *
## s(icu)                      4.990  6.271  3.268 0.00323 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.956   Deviance explained = 97.7%
## -ML = 781.31  Scale est. = 1         n = 460
```

```
plot.gam(mod_refit, shade = TRUE, pages = 1)
```



We first look at anova of the previous fit model. The predictors *people_fully_vaccinated* and *tests* have edf = 1, which implies that the terms can be explained without having to smooth them. Hence we first remove smooths and refit the model.

```
anova.gam(mod_refit)
```

```
##
## Family: Negative Binomial(34.636)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + s(tests) + s(vaccines) +
##     s(people_fully_vaccinated) + s(hosp) + s(icu)
##
## Approximate significance of smooth terms:
##                               edf Ref.df      F p-value
## s(Day)                     24.961 26.785 24.287 < 2e-16
## s(tests)                    1.000  1.000 65.803 < 2e-16
## s(vaccines)                 7.997  8.460  5.746   1e-06
## s(people_fully_vaccinated)  1.000  1.000  3.191 0.07476
## s(hosp)                     5.017  6.182  2.664 0.02812
## s(icu)                      4.990  6.271  3.268 0.00323
```

```
refit1 <- bam(daily_confirmed_cases ~ s(Day, k=30)+tests+s(vaccines)+people_f
ully_vaccinated+s(hosp)+s(icu), data = data, family = nb(link = "log"), metho
d = "ML")
```

```
summary.gam(refit1)
```

```
##
## Family: Negative Binomial(29.068)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + s(vaccines) +
##     (people_fully_vaccinated) + s(hosp) + s(icu)
##
## Parametric coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -46.239     10.228  -4.521 8.00e-06 ***
## tests                      6.575      1.332   4.938 1.14e-06 ***
## people_fully_vaccinated    3.462      1.379   2.510   0.0124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df      F  p-value
## s(Day)      27.53 28.705 83.189  < 2e-16 ***
## s(vaccines)  1.00  1.000  0.846    0.358
## s(hosp)      1.00  1.000  0.008    0.929
```

```
## s(icu)       3.92  5.059  6.336 1.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.947   Deviance explained = 97.2%
## -ML = 745.76  Scale est. = 1         n = 460
```

We now observe that the predictors *vaccines* and *hosp* hold no significance with the model and their edf =1, which implies they can be modelled as in linear relationship with the response. Hence, we remove smooth for these predictors and refit the model before removing any variables from the model. We now refit the new model.

```
anova.gam(refit1, test = "Chisq")
```

```
##
## Family: Negative Binomial(29.068)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + s(vaccines) +
##     (people_fully_vaccinated) + s(hosp) + s(icu)
##
## Parametric Terms:
##                          df      F  p-value
## tests                     1 24.380 1.14e-06
## people_fully_vaccinated  1  6.301   0.0124
##
## Approximate significance of smooth terms:
##                edf Ref.df      F  p-value
## s(Day)      27.532 28.705 83.189  < 2e-16
## s(vaccines)  1.000  1.000  0.846    0.358
## s(hosp)      1.000  1.000  0.008    0.929
## s(icu)       3.920  5.059  6.336 1.03e-05
```

```
# model after removing smooths for a few predictors
refit2 <- bam(daily_confirmed_cases ~ s(Day, k=30)+(tests)+(vaccines)+(people
_fully_vaccinated)+(hosp)+s(icu), data = data, family = nb(link = "log"), met
hod = "ML")
```

```
summary.gam(refit2)
```

```
##
## Family: Negative Binomial(29.068)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (vaccines) +
##     (people_fully_vaccinated) + (hosp) + s(icu)
##
## Parametric coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)             -4.070e+01  9.109e+00  -4.469 1.01e-05 ***
## tests                    6.575e+00  1.332e+00   4.938 1.14e-06 ***
## vaccines                -1.125e+00  1.223e+00  -0.920   0.3581
## people_fully_vaccinated  3.462e+00  1.379e+00   2.510   0.0124 *
## hosp                    -2.194e-05  2.459e-04  -0.089   0.9289
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(Day) 27.532 28.705 83.193  < 2e-16 ***
## s(icu)  3.919  5.058  6.337 1.03e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.947   Deviance explained = 97.2%
## -ML = 745.75  Scale est. = 1         n = 460
```

We now observe that the predictors *vaccines* and *hosp* hold no significance with the model. Both their p-values are greater than 0.01, Hence, we conclude that these predictors are not significant to the model at 90% confidence level and remove them and refit the model.

```
anova.gam(refit2, test = "Chisq")
```

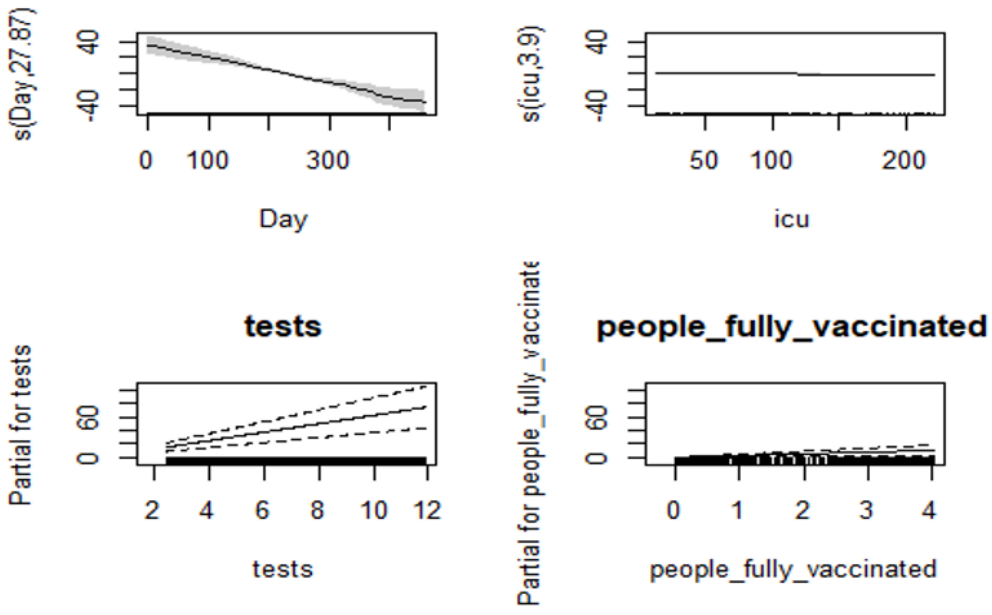
```
##
## Family: Negative Binomial(29.068)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (vaccines) +
##     (people_fully_vaccinated) + (hosp) + s(icu)
##
## Parametric Terms:
##                          df      F  p-value
## tests                     1 24.379 1.14e-06
## vaccines                  1  0.846   0.3581
## people_fully_vaccinated   1  6.301   0.0124
## hosp                      1  0.008   0.9289
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(Day) 27.532 28.705 83.193  < 2e-16
## s(icu)  3.919  5.058  6.337 1.03e-05
```

```
# remove insignificant predictors
refit3 <- bam(daily_confirmed_cases ~ s(Day, k=30)+(tests)+(people_fully_vacc
inated)+s(icu), data = data, family = nb(link = "log"), method = "ML")

summary.gam(refit3)
```

```
##
## Family: Negative Binomial(28.987)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_vaccinate
d) + s(icu)
##
## Parametric coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -41.733      9.030  -4.622 5.05e-06 ***
## tests                      6.227      1.282   4.856 1.69e-06 ***
## people_fully_vaccinated    2.576      0.977   2.637  0.00868 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df       F p-value
## s(Day) 27.866 28.819 131.119  <2e-16 ***
## s(icu)  3.895  5.025   8.154  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.947   Deviance explained = 97.2%
## -ML = 745.19  Scale est. = 1         n = 460
```

The above summary indicates, that removal of the insignificant terms did not make a difference to the model residuals as indicated by R^2 and ML values remains unchanged at 745(decreases a few decimal points since the previous fit indicating a better model)

```
plot.gam(refit3, shade = TRUE, all.terms = TRUE, pages = 1)
```



Question 11

Fit Categorical Variables

We fit each categorical variable one-by-one and compare the model to the previous model to obtain it's significance using anova() tests.

```
# Fit the model with school_closing predictor
mod_c1 <- bam(daily_confirmed_cases ~ s(Day, k=30)+(tests)+(people_fully_vacc
inated)+s(icu)+school_closing, data = data, family = nb(link = "log"), method
= "ML")
summary.gam(mod_c1)

##
## Family: Negative Binomial(29.384)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_vaccinate
d) +
##     s(icu) + school_closing
##
## Parametric coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -42.81941    9.01251  -4.751 2.78e-06 ***
## tests                     6.39666    1.27958   4.999 8.46e-07 ***
## people_fully_vaccinated   2.53256    0.97504   2.597  0.00972 **
## school_closing1          -0.02162    0.14620  -0.148  0.88250
## school_closing2          -0.07706    0.21321  -0.361  0.71797
## school_closing3           0.32209    0.27291   1.180  0.23859
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F p-value
## s(Day) 27.779 28.810 83.827  <2e-16 ***
## s(icu)  3.713  4.795  9.935  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.947   Deviance explained = 97.2%
## -ML =  745.6  Scale est. = 1         n = 460

anova.gam(refit3, mod_c1, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu)
```

```
## Model 2: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu) + school_closing
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    421.51     6573.6
## 2    419.33     6567.3 2.1802   6.2649  0.05177 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above anova test, we can say that the addition of school_closing predictor holds significance to us at 90% confidence since the p-value of the test is < 0.1

```
# Fit the model with gathering_restrictions predictor
mod_c2 <- bam(daily_confirmed_cases ~ s(Day, k=30)+(tests)+(people_fully_vacc
inated)+s(icu)+gatherings_restrictions , data = data, family = nb(link = "log
"), method = "ML")
summary.gam(mod_c2)

##
## Family: Negative Binomial(29.824)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_vaccinate
d) +
##     s(icu) + gatherings_restrictions
##
## Parametric coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -50.8920     9.6096  -5.296 1.91e-07 ***
## tests                       7.5344     1.3689   5.504 6.46e-08 ***
## people_fully_vaccinated     2.6997     0.9782   2.760  0.00603 **
## gatherings_restrictions1   -0.2377     0.2355  -1.009  0.31347
## gatherings_restrictions2   -0.3260     0.2367  -1.377  0.16918
## gatherings_restrictions3   -0.1128     0.1875  -0.602  0.54775
## gatherings_restrictions4   -0.2799     0.1571  -1.782  0.07548 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df       F  p-value
## s(Day) 27.942 28.847 103.294  < 2e-16 ***
## s(icu)  3.763  4.872   7.531 2.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.952   Deviance explained = 97.3%
## -ML =  748.2  Scale est. = 1         n = 460

anova.gam(refit3, mod_c2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu)
## Model 2: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu) + gatherings_restrictions
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    421.51     6573.6
## 2    417.92     6560.8 3.5906   12.801 0.008761 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The above anova test implies that the addition of gathering_restrictions predictor holds significance to us at 99% confidence since the p-value of the test is < 0.01

```
# Fit the model with transport_closing predictor
mod_c3 <- bam(daily_confirmed_cases ~ s(Day, k=30)+(tests)+(people_fully_vacc
inated)+s(icu)+transport_closing , data = data, family = nb(link = "log"), me
thod = "ML")
summary.gam(mod_c3)

##
## Family: Negative Binomial(29.092)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_vaccinate
d) +
##     s(icu) + transport_closing
##
## Parametric coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -41.2144     9.0215  -4.568 6.45e-06 ***
## tests                     6.1453     1.2817   4.795 2.26e-06 ***
## people_fully_vaccinated   2.5599     0.9752   2.625  0.00898 **
## transport_closing1        0.2168     0.1501   1.444  0.14953
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df       F  p-value
## s(Day) 27.833 28.812 119.087  < 2e-16 ***
## s(icu)  3.772  4.877   8.493 2.74e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.947   Deviance explained = 97.2%
## -ML =  745.3  Scale est. = 1         n = 460
```

```
anova.gam(refit3, mod_c3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu)
## Model 2: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu) + transport_closing
##   Resid. Df Resid. Dev      Df Deviance Pr(>Chi)
## 1    421.51     6573.6
## 2    420.59     6571.9 0.91691   1.6807   0.1759
```

The above anova test indicates that the addition of transport_closing predictor proves to be an insignificant addition to the model since the p-value of the test is > 0.1, i.e., we are not confident at any significance level that the variable would make a difference to the models predictions.

We now add the 2 significant categorical variables to the model and fit the model using gam()

```
mod_c4 <- bam(daily_confirmed_cases ~ s(Day, k=30)+(tests)+(people_fully_vacc
inated)+s(icu)+school_closing+gatherings_restrictions , data = data, family =
nb(link = "log"), method = "ML")
summary.gam(mod_c4)
```

```
##
## Family: Negative Binomial(30.278)
## Link function: log
##
## Formula:
## daily_confirmed_cases ~ s(Day, k = 30)+(tests)+(people_fully_vaccinated) +
##     s(icu) + school_closing + gatherings_restrictions
##
## Parametric coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -51.79730    9.58232  -5.406 1.09e-07 ***
## tests                      7.68271    1.36526   5.627 3.36e-08 ***
## people_fully_vaccinated    2.64359    0.97574   2.709  0.00702 **
## school_closing1           -0.02932    0.14638  -0.200  0.84135
## school_closing2           -0.06610    0.21345  -0.310  0.75696
## school_closing3            0.33318    0.27282   1.221  0.22268
## gatherings_restrictions1  -0.25292    0.23493  -1.077  0.28229
## gatherings_restrictions2  -0.32663    0.23621  -1.383  0.16745
## gatherings_restrictions3  -0.11707    0.18823  -0.622  0.53431
## gatherings_restrictions4  -0.28038    0.15752  -1.780  0.07581 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F p-value
## s(Day) 27.858  28.84 64.687  <2e-16 ***
```

```
## s(icu)  3.771   4.87  8.992  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.952   Deviance explained = 97.3%
## -ML = 748.95  Scale est. = 1         n = 460
```

```
anova.gam(refit3, mod_c4, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu)
## Model 2: daily_confirmed_cases ~ s(Day, k = 30) + (tests) + (people_fully_
vaccinated) +
##     s(icu) + school_closing + gatherings_restrictions
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    421.51     6573.6
## 2    415.30     6553.8 6.2101   19.735 0.003597 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# check for overdispersion - close to 1
sum(residuals(mod_c4, type = "pearson")^2) / df.residual(mod_c4)
```

```
## [1] 1.047754
```

We can now say from the hypothesis test that addition of these 2 variables holds significance to the model. But we can also observe that some variable categories don't hold too much significance to the response variable. We can also see a decrease in residual deviance.

We also check for over-dispersion of the final model. The output of residuals/no. of df is very close to 1 indicating the model has no over or under dispersion.

Question 12

We can describe the model fit by analyzing the R^2 results from the model summary. R^2 indicates the fraction of model explained by the predictor variables.

R^2 of our model is 0.952 and deviance explained is 97.2%, i.e., almost 95% of our model is explained by the predictors which implies a good model.

Question 13

The Generalized Additive Models do follow all the assumptions of Generalized Linear model, but it relaxes the assumption that each of the independent predictors have to vary linearly with the dependent response variable.

Assumptions :

1. The data Y1,Y2,…,Yn are independently distributed, i.e., cases are independent.
2. The dependent variable Yi does NOT need to be normally distributed, but it typically assumes a distribution from an exponential family (e.g. binomial, Poisson, multinomial, normal, etc.).
3. GAMs assume that the observations are i.i.d(Identical independent distribution)
4. Residuals need to be independent.
5. Residuals are normally distributed $\in_i \sim N(0,1)$
6. Standardized residuals should have approximately equal/constant variance

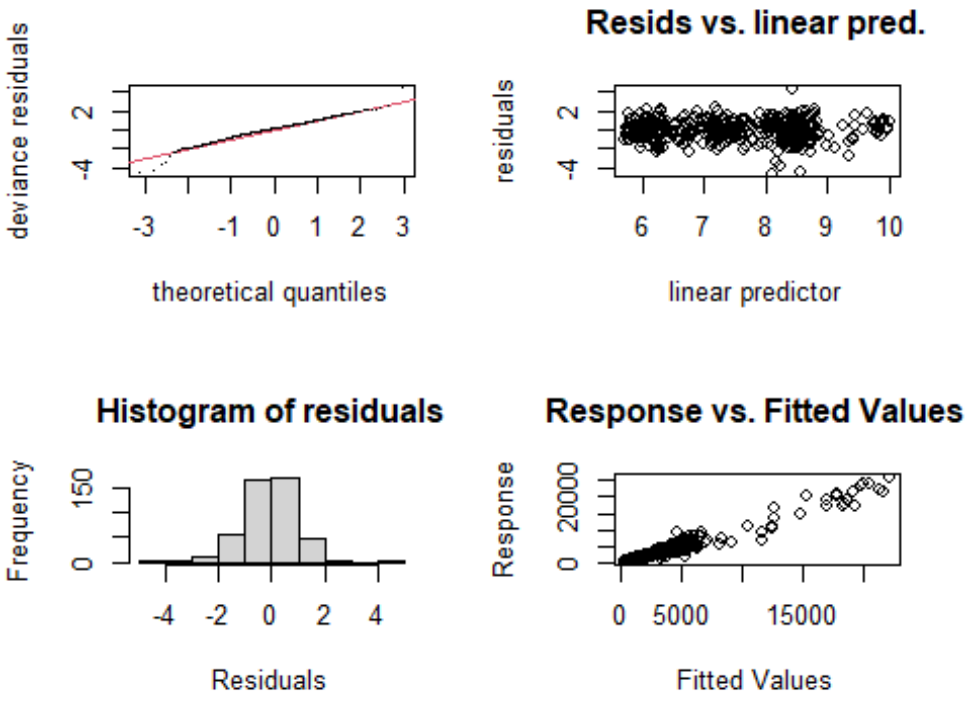
Question 14

We can say that the data collected about each day does not relate to the previous day Covid-19 cases.

The below residual QQ-plot indicate good fit of data and residual vs fitted plot clearly shows the zero mean and constant variance by the points distributed almost equally at y=0.

The histogram indicates the normal distribution of data.

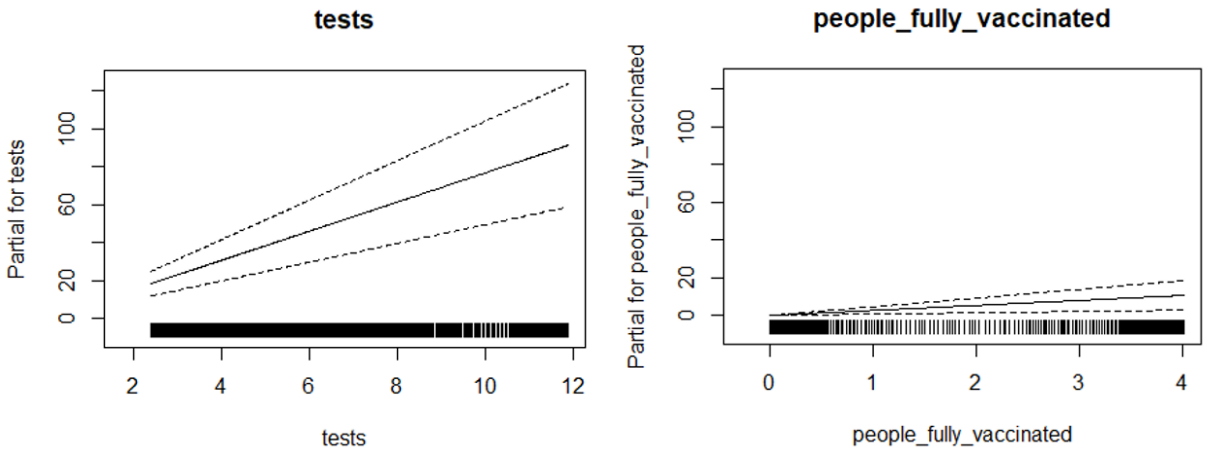
```
gam.check(mod_c4)
```




```
dwt(mod_c4)

##  lag Autocorrelation D-W Statistic p-value
##    1       0.1044321      1.765192       0
##  Alternative hypothesis: rho != 0
```

The Durbin Watson test statistic is used to check for random (or independent identically distributed) sample. The p-value for the statistic is higher than the threshold (0.05) so the hypothesis of no correlation among residuals is rejected and as the test is within an acceptable range of 1.5 to 2.5 this is not cause for concern.

Question 15



Interpretation :

Test predictor varies linearly with the log of the average number of Covid-19 cases per day, i.e., has a positive effect on response variable. As the number of tests conducted per day increases by 1, the average number covid cases increases by Exp(7.68271)≈2170 cases.

People_fully_vaccinated predictor varies linearly with the log of the average number of Covid-19 cases per day, i.e., has a very slight positive effect on response variable. As the number of fully vaccinated people per day increases by 1, the average number covid cases increases by Exp(2.64359)≈14 cases(This might indicate that the vaccination is not much of an effect if the whole population is not vaccinated and follow certain rules).

Keeping number of people fully vaccinated and number of tests conducted constant, if there are no restrictions applied by the government, the number of covid-19 cases increases by exp(-51.79730)≈ negligible. We can also observe that there is no significance for the other groups of school_gathering restrictions.

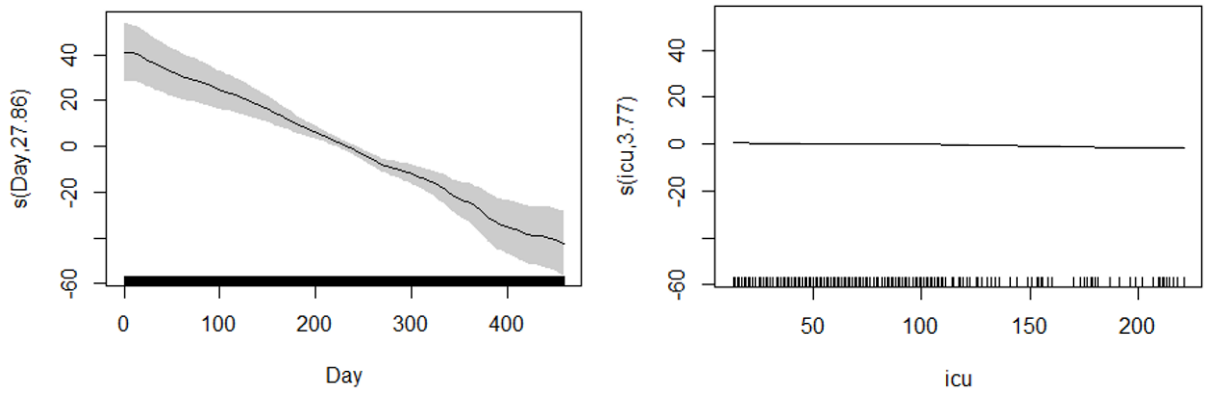
When full restrictions is compared to no gathering restrictions, the log average value od number of Covid-19 cases, decreases approximately by -51.5

Question 16

From the above plot we can see that as the number of days increases, the log of average number of Covid-19 cases decreases steeply. It is a significant variable

We can observe that when the count of icu patients increase by 1, the count of covid 19 also increases but by a very less margin.

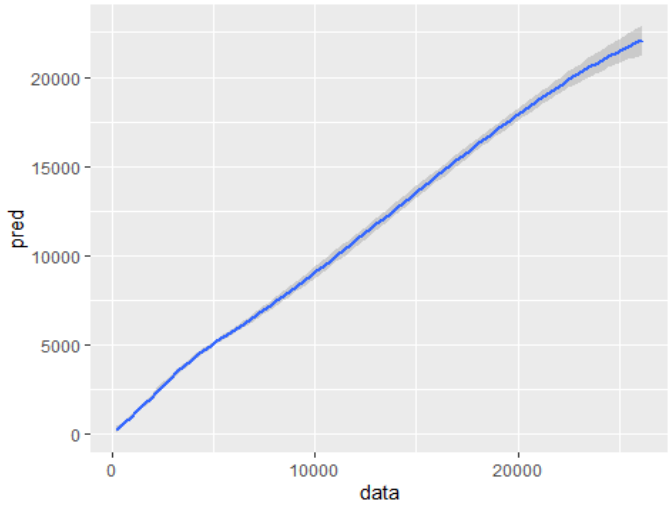
```
plot(mod_c4, shade = TRUE, all.terms = TRUE)
```



Question 17

Prediction :

```
prediction <- predict.bam(mod_c4, type="response", se=TRUE)

p <- data.frame("data" = data$daily_confirmed_cases, "pred" = prediction$fit)
ggplot(p, aes(data, pred), title(main = "Prediction plot"))+geom_smooth()

mean(prediction$se.fit)

## 166.0269
```



We can observe from the prediction plot that the data fits perfectly well with very less standard error of 166 for a huge population. From the deviance residuals from the summary, we can say

that the model accounts for almost 97% of the variance in response. Hence, the model is concluded to be a very good fit for the data.

Question 18

As from the model's output, we can observe that the categorical variables do not form a significant variable in the model. The group variables are not significant enough to conclude the effects of government restrictions on the output.

We can say that from the model fit that there is negligible effect of government restrictions on the count of Covid-19 cases per day. This might be due to the lack of consideration of continuous data and past effects on the model. The categorical variables vary linearly and modelling them with log function affects the significance and interferes in the inference of the effects of government restrictions on the number of Covid-19 cases per day. As seen from the plot earlier, the categorical variables vary almost linearly with the response and outcomes are observed as expected. With complete restrictions the number of Covid-19 cases per day must ideally decrease. But due to the model not being linear, it affects the interpretation of the government restrictions.