## Table of contents

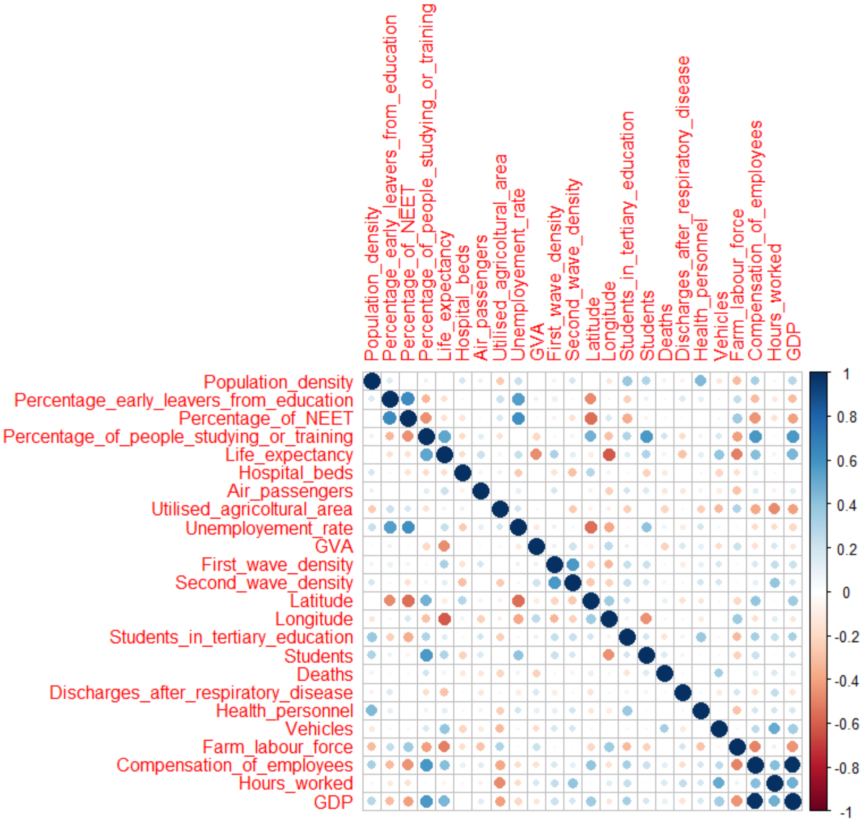
## Data preprocessing

1. Filling missing data from Eurostat
2. Changing names of columns
3. Scaling features by population (amount per 100.000 inhabitants)
    a. STUDENTS_IN_TERTIARY_EDUCATION
    b. STUDENTS
    c. DEATHS
    d. DISCHARGES_AFTER_RESPIRATORY_DISEASE
    e. HEALTH_PERSONNEL
    f. VEHICLES
    g. FARM_LABOUR_FORCE
    h. COMPENSATION_OF_EMPLOYEES
    i. HOURS_WORKED
    j. GDP
4. Dropping population feature

## Data exploration

**Correlation matrix**

As it can be observed on the picture below, there are plenty of features strongly correlated with each other which might introduce some kind redundancy into our models. It might be worth considering dropping some features in order to prevent models from overfitting and ensure better generalization abilities.

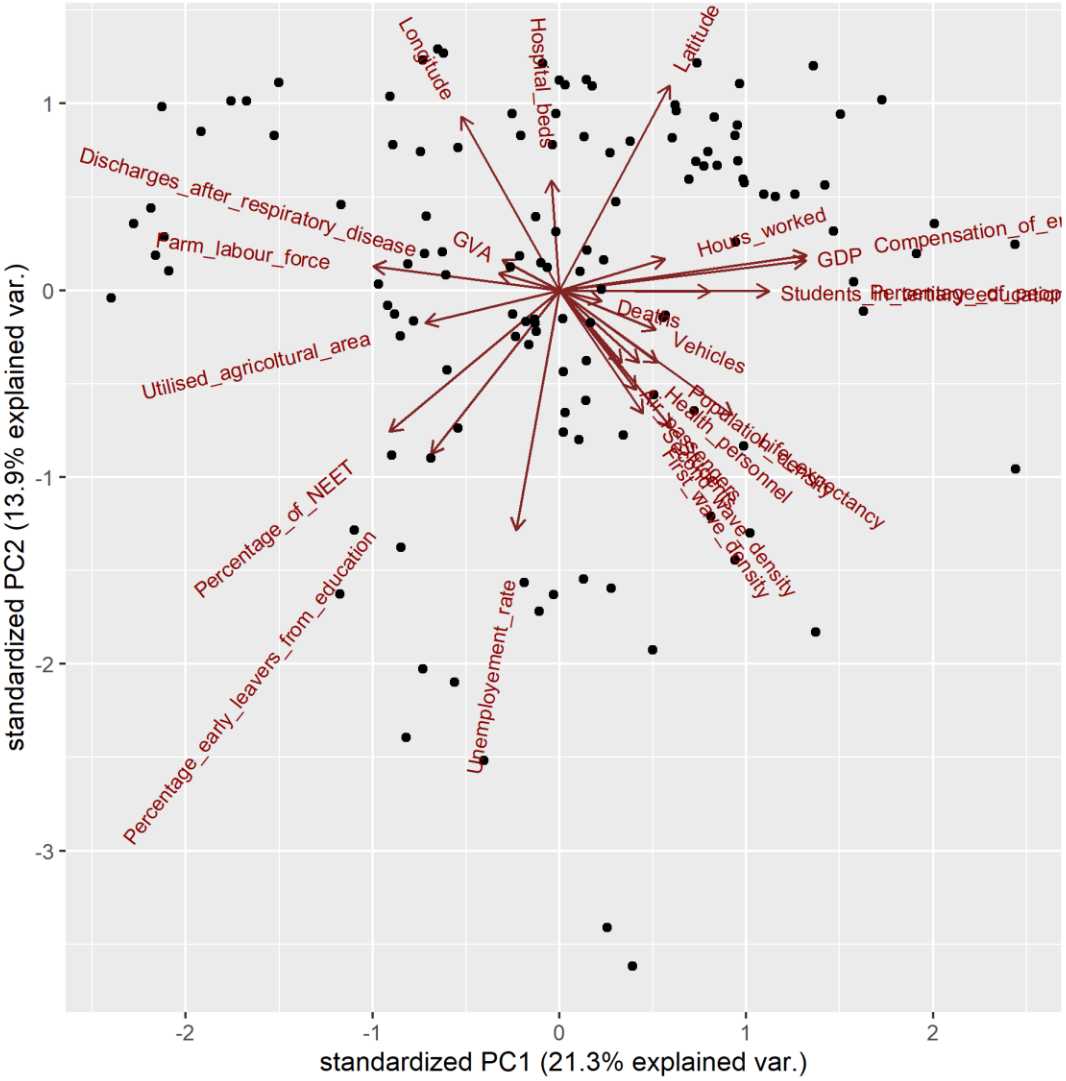


Strong correlations:

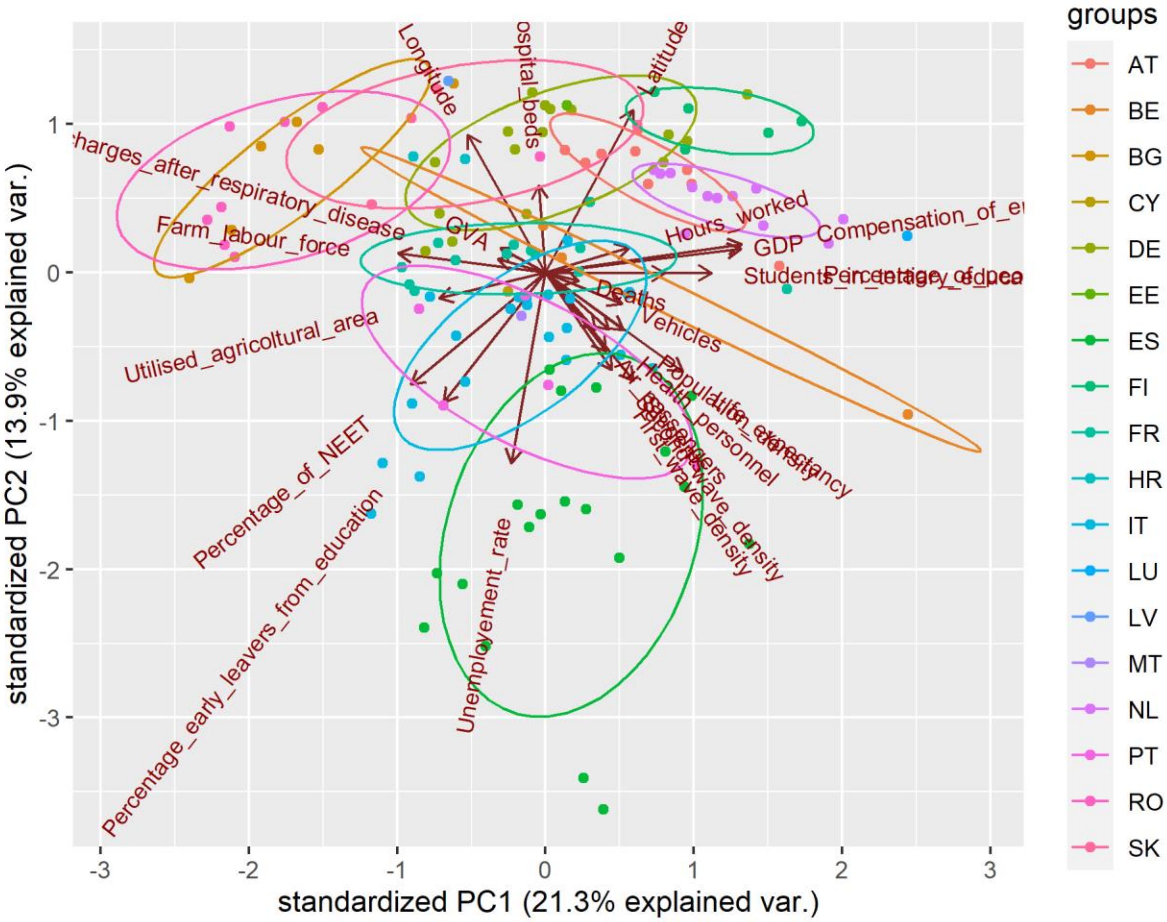
- GDP AND COMPENSATION_OF_EMPLOYEES
- GDP AND PERCENTAGE_OF_PEOPLE_STUDYING_OR_TRAINING
- COMPENSATION_OF_EMPLOYEES AND PERCENTAGE_OF_PEOPLE_STUDYING_OR_TRAINING
- UNEPLOYMENT_RATE AND PERCENTAGE_OF_NEET
- UNEPLOYMENT_RATE AND PERCENTAGE_OF_EARLY_LEAVERS_FROM_EDUCATION
- PERCENTAGE_OF_NEET AND PERCENTAGE_OF_EARLY_LEAVERS_FROM_EDUCATION
- STUDENTS AND PERCENTAGE_OF_PEOPLE_STUDYING_OR_TRAINING (WHICH IS OBVIOUS)

No strong correlations between any variable and density of cases during first and second wave have been observed.

## PCA

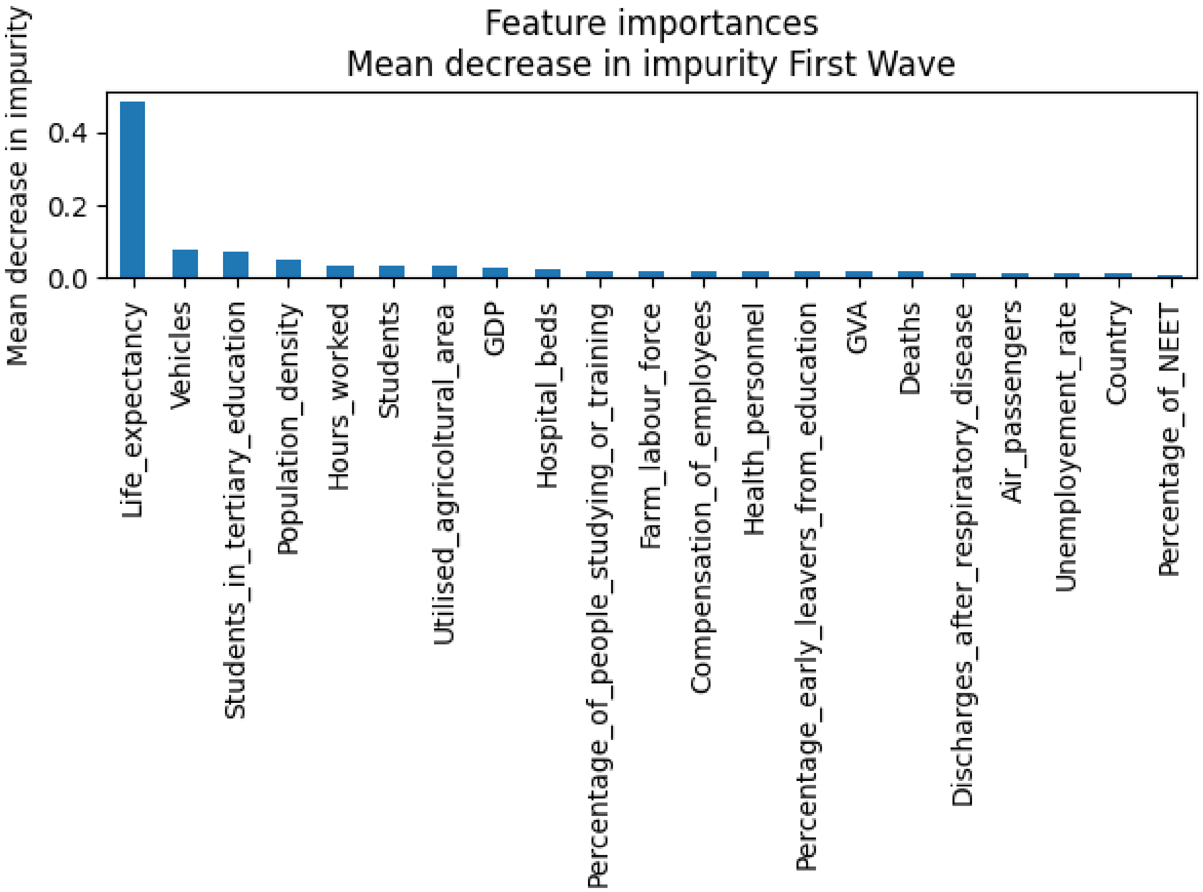


Visualization of PCA also gives a valuable insight into characteristics of the dataset. For example we can clearly see that regions with high values of **AIR_PASSENGERS**, **LIFE_EXPTECTANCY, POPULATION_DENSITY, VEHICLES** and **HEALTH_PERSONNEL** tend to have high density of cases during the first and second wave, while these with high values of **GVA, HOSPITAL_BEDS** and **DISCHARGES_AFTER_RESPIRATORY_DISEASE** had less infections. All these observations seem to have a reasonable, intuitive explanation.

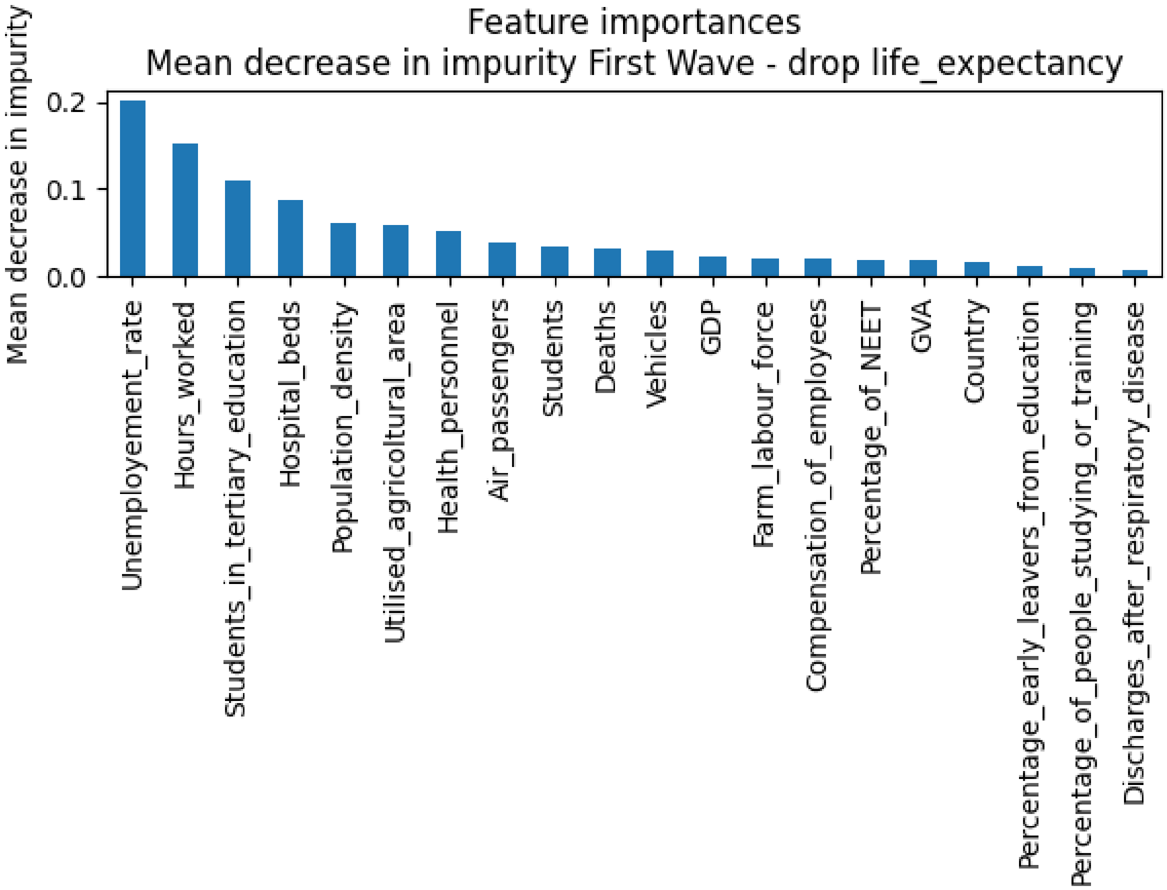


Moreover, grouping of observations by countries highlights that observations from the same countries tend to be close to each other in the features space. It might be a suggestion to use **linear mixed effect** model to capture hierarchical characteristic of the dataset.

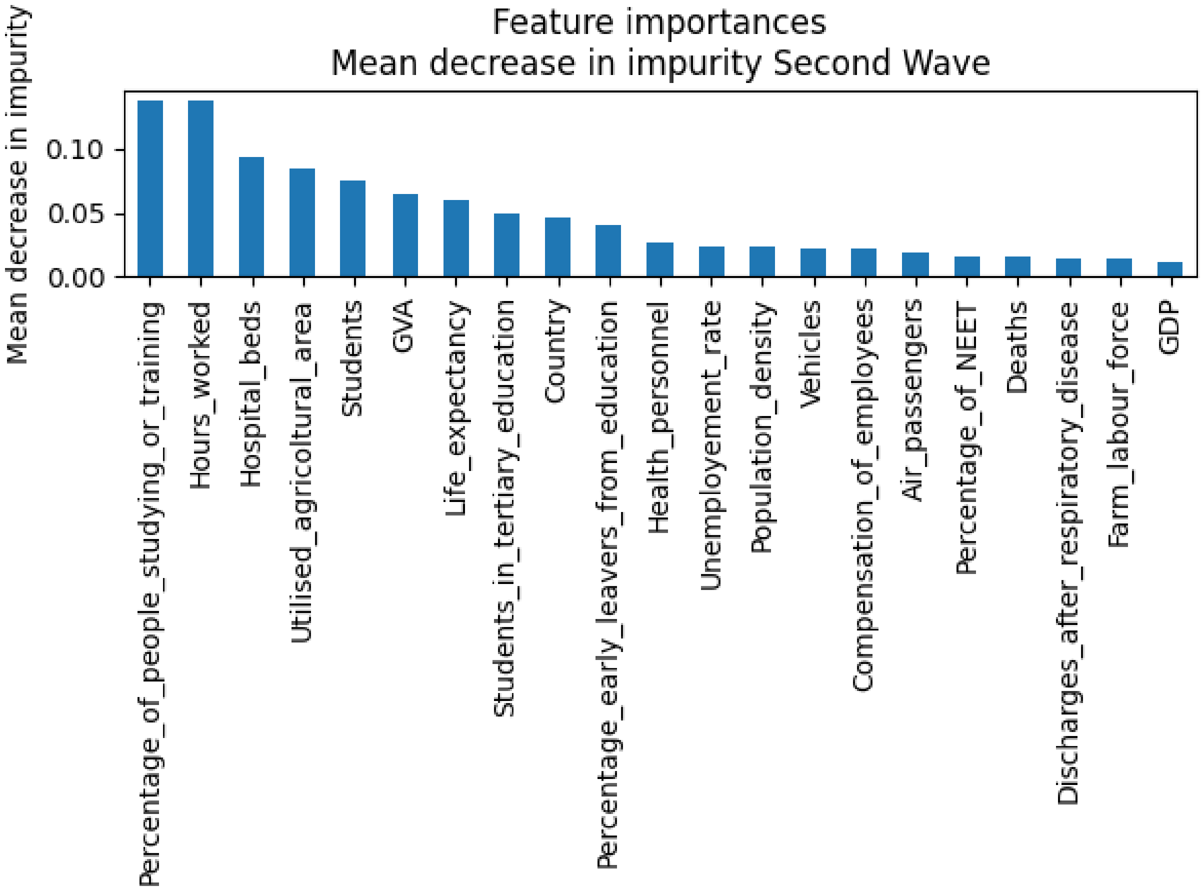
**Random forest**



In order to predict density of cases during first wave it seems that decision of our model is mainly based on **LIFE_EXPECTANCY** feature. It is, however, hard to distinguish another valuable features so in the next trial life_expectancy feature was deleted from the dataset and model was trained again.



This experiment revealed other relevant features with high predictive power like **UNEMPLOYMENT_RATE, HOURS_WORKED, STUDENTS_IN_TERTIARY_EDUCATION, HOSPITAL_BEDS, POPULATION_DENSITY.**

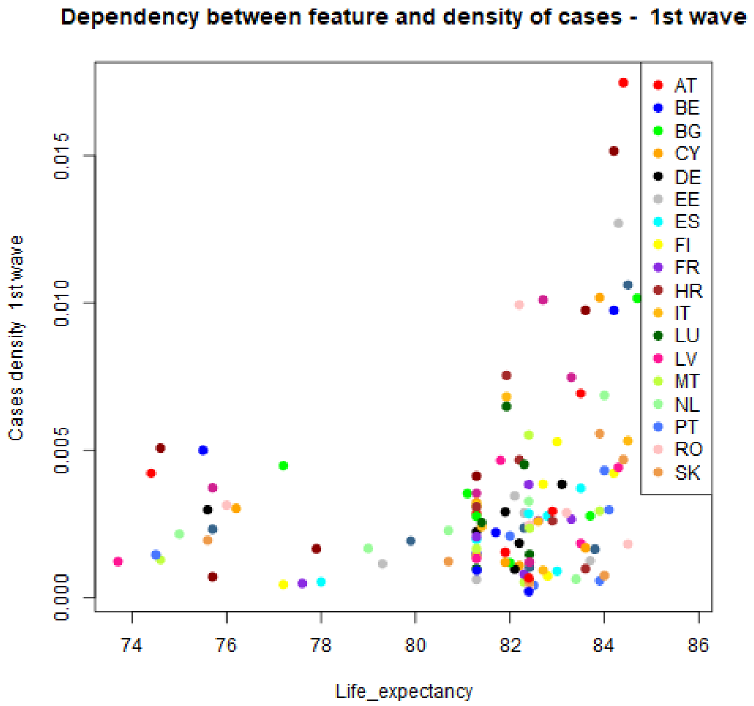


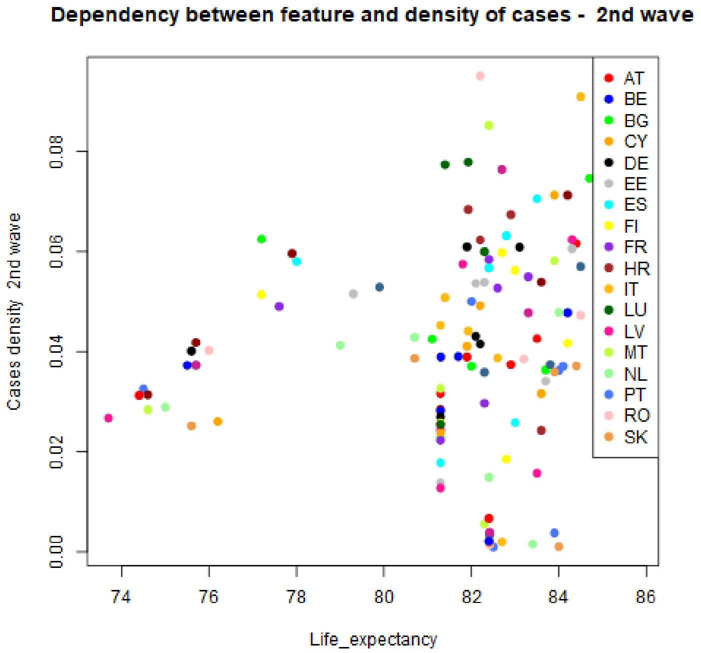
In case of the second wave the most significant features seem to be **PERCENTAGE_OF_PEOPLE_STUDYING_OR_TRAINING, HOURS_WORKED, HOSPITAL_BEDS, UTILISED_AGRICULTURAL_AREA, STUDENTS, GVA, LIFE_EXPECTANCY.**

It is pretty interesting that country wasn't taken into account that much in neither case, which is opposite of our predictions that regions from the same country might be strongly correlated due to the same applied policies and so on.

Furthermore, life_expectancy is not a key factor anymore for the second wave, but rather professional activity is crucial.

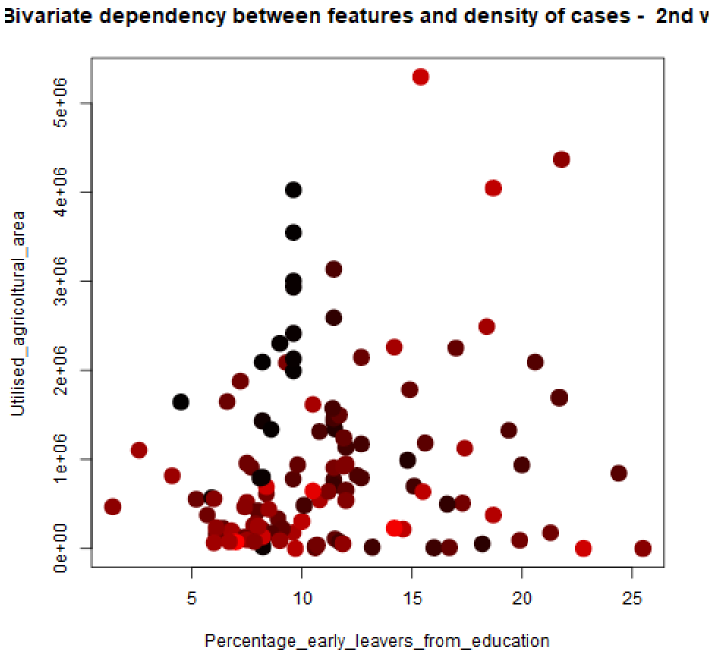
## Features against cases density

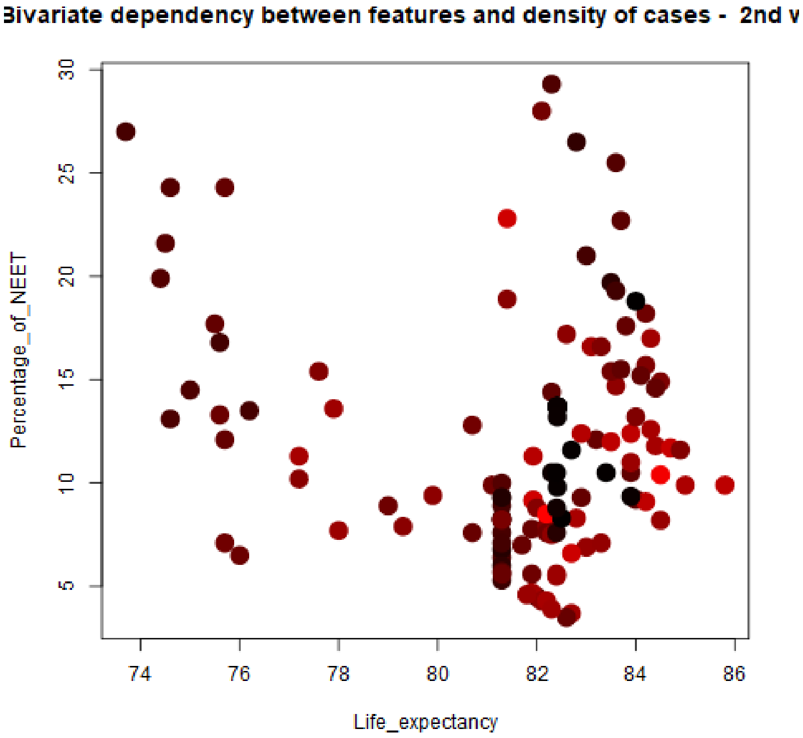




Analysis of all the features didn’t reveal any more interesting dependencies or trends. For **LIFE_EXPECTANCY** it can be observed that high values of cases occur only for regions with high life expectancy but there is no clear trend apart from it.

## Bivariate analysis

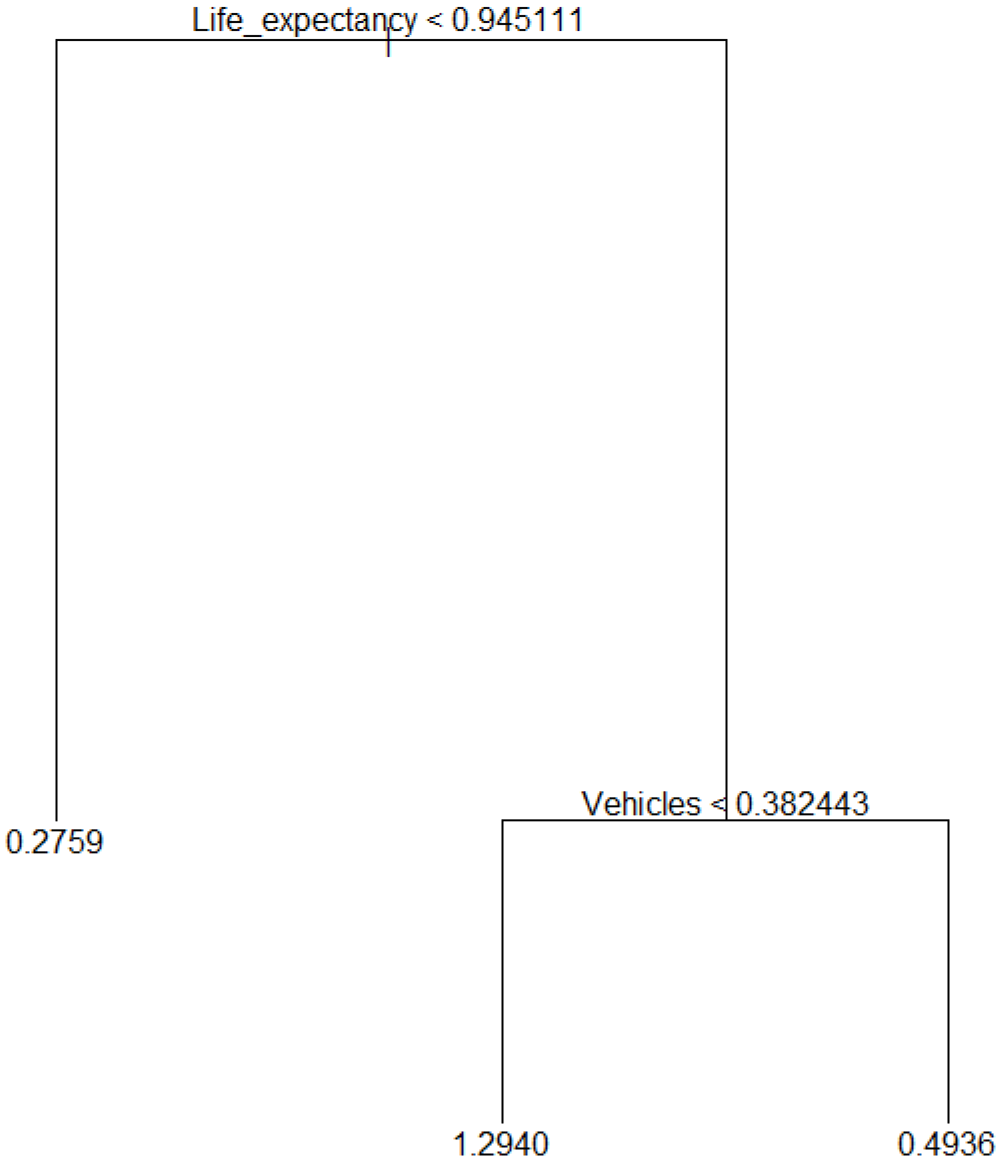




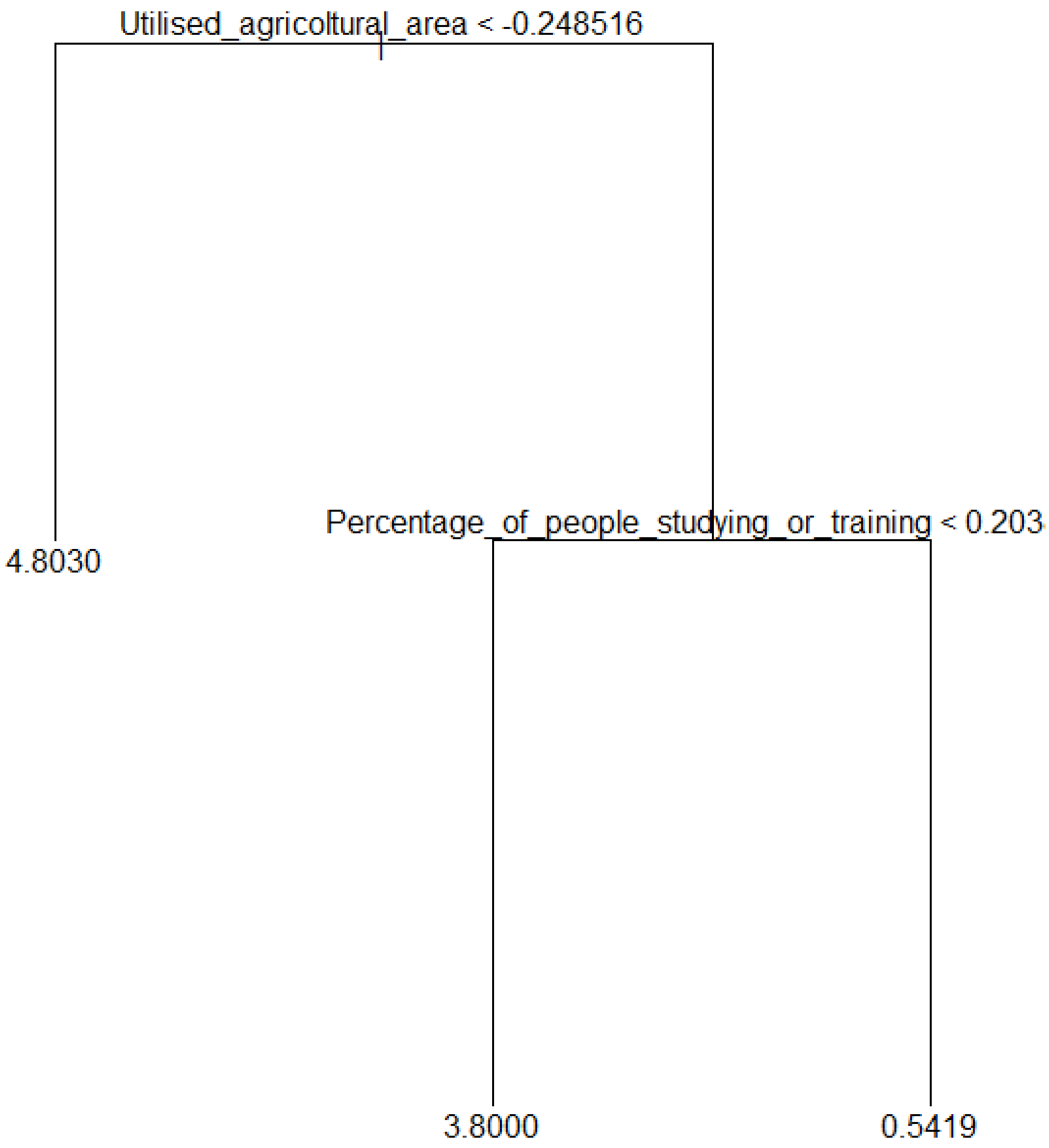
Bivariate analysis helps to find some interactions between features, however, it seems to be hard to model any valuable interactions which could be used in linear regression model.

## Regression tree

**First wave**

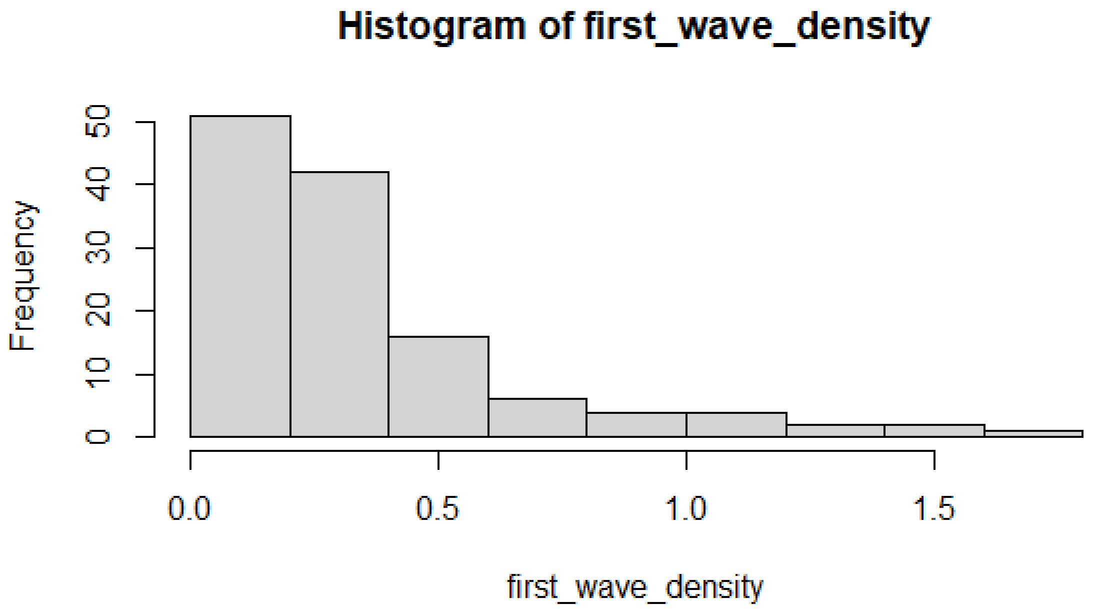

Life_expectancy < 0.945111
0.2759
Vehicles < 0.382443
1.2940
0.4936

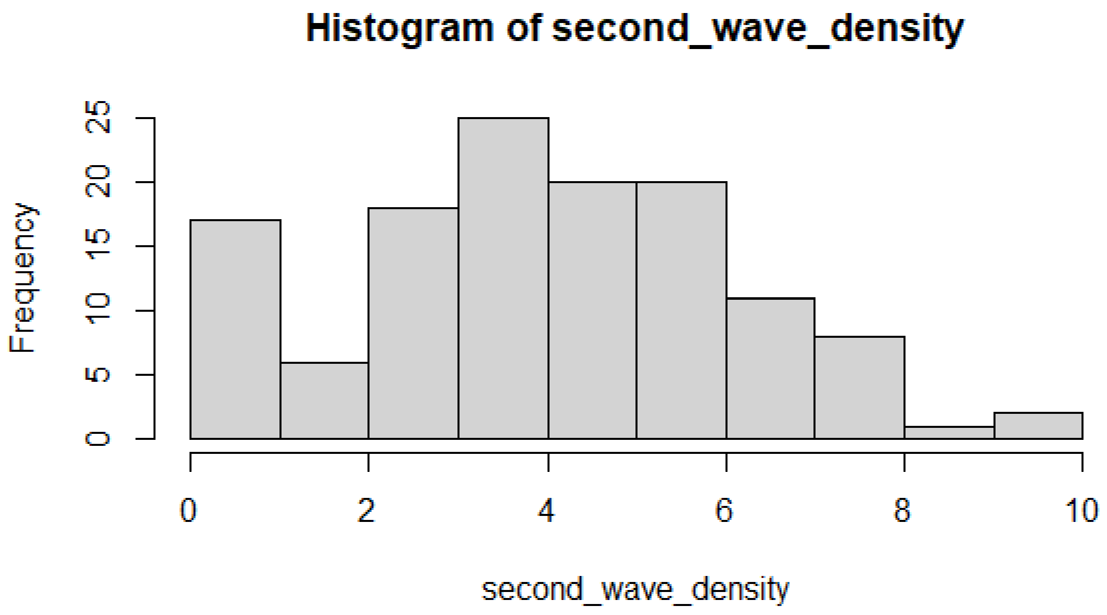
**Second wave**


Utilised_agricoltural_area < -0.248516
4.8030
Percentage_of_people_studying_or_training < 0.203
3.8000
0.5419

Professor, mentioned that regression trees can be used to observe some interactions between features and model them into the final model, however, I don't know how these could be modelled. Trees were pruned, according to a cross validation results.

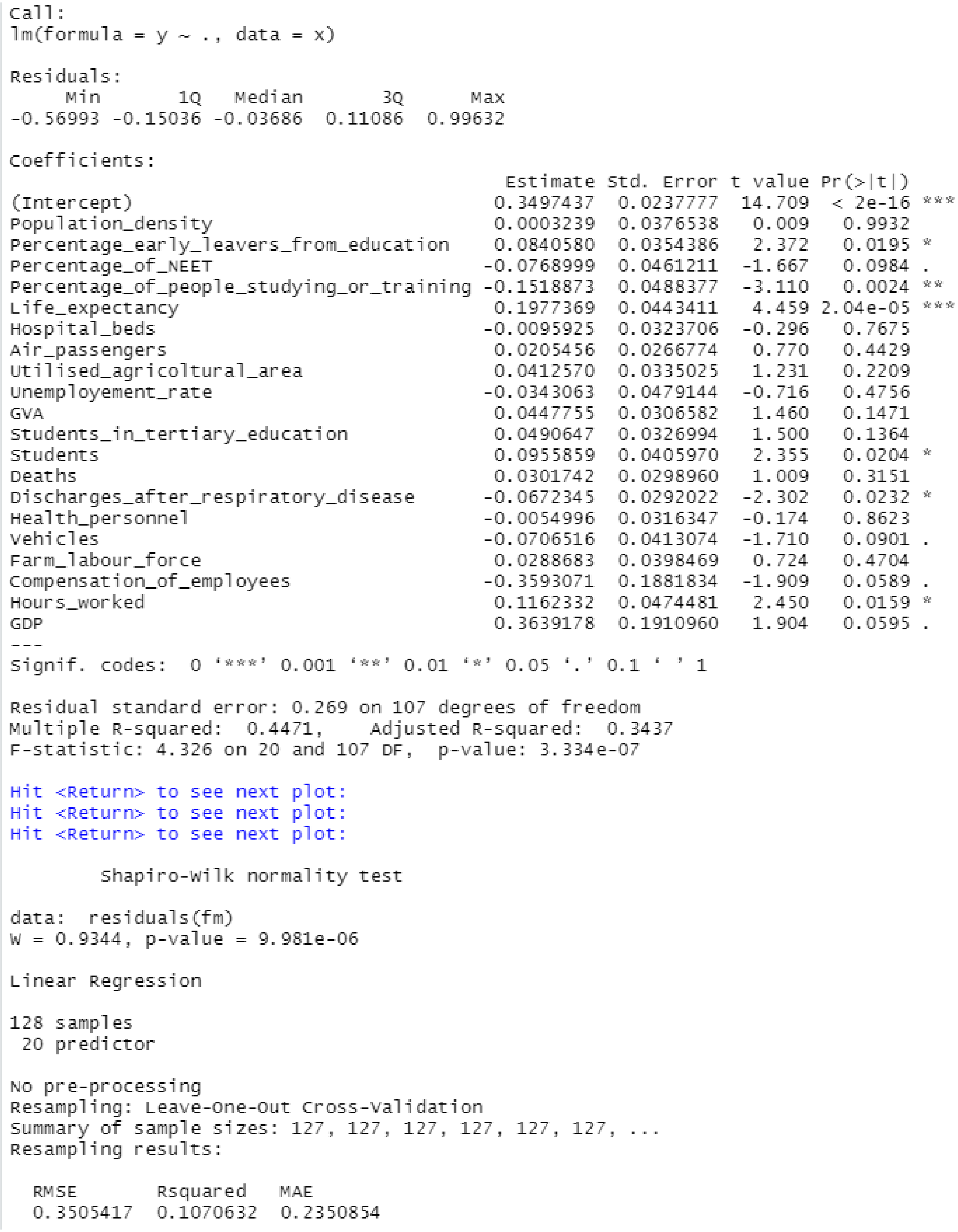
### Linear regression



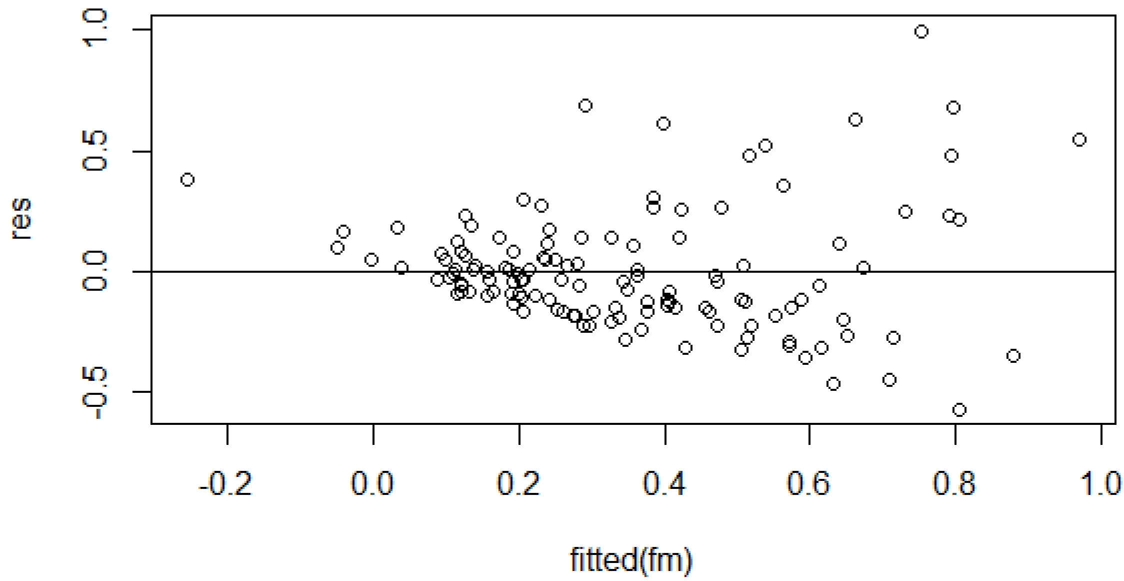


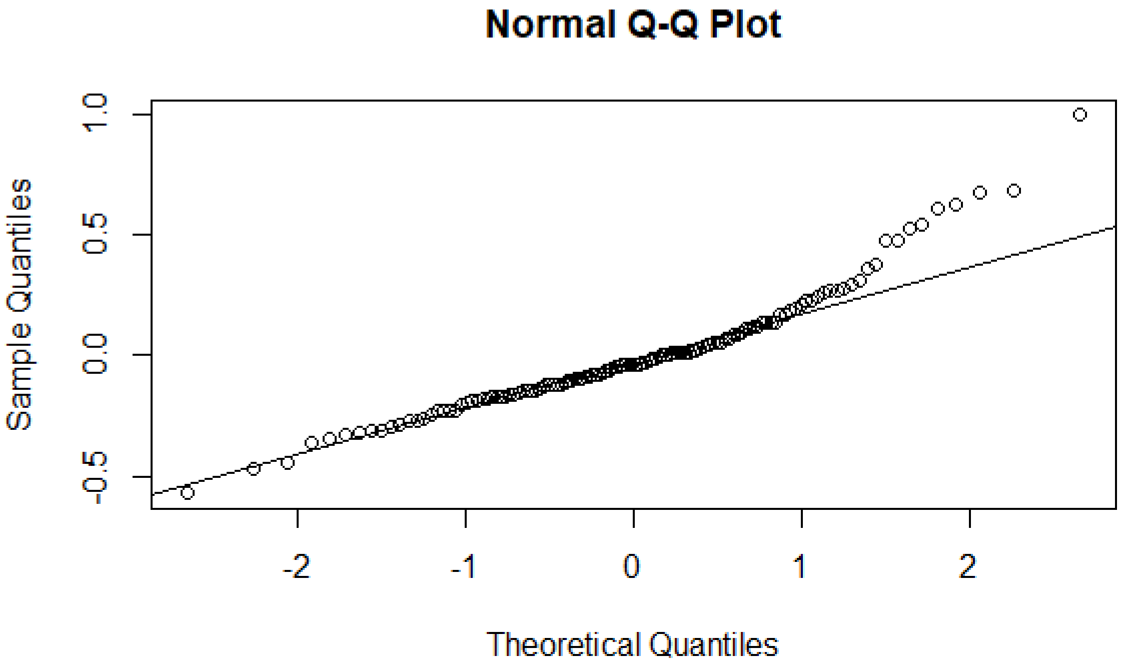
Distributions of cases was demonstrated just in order to give a brief insight on relatively how big mistakes our predictors are doing. Furthermore, these values were rescaled to represent number of cases per 10.000.000 inhabitants instead of 100.000, in order to make them easier to comprehend.

## Linear regression with all features - First wave

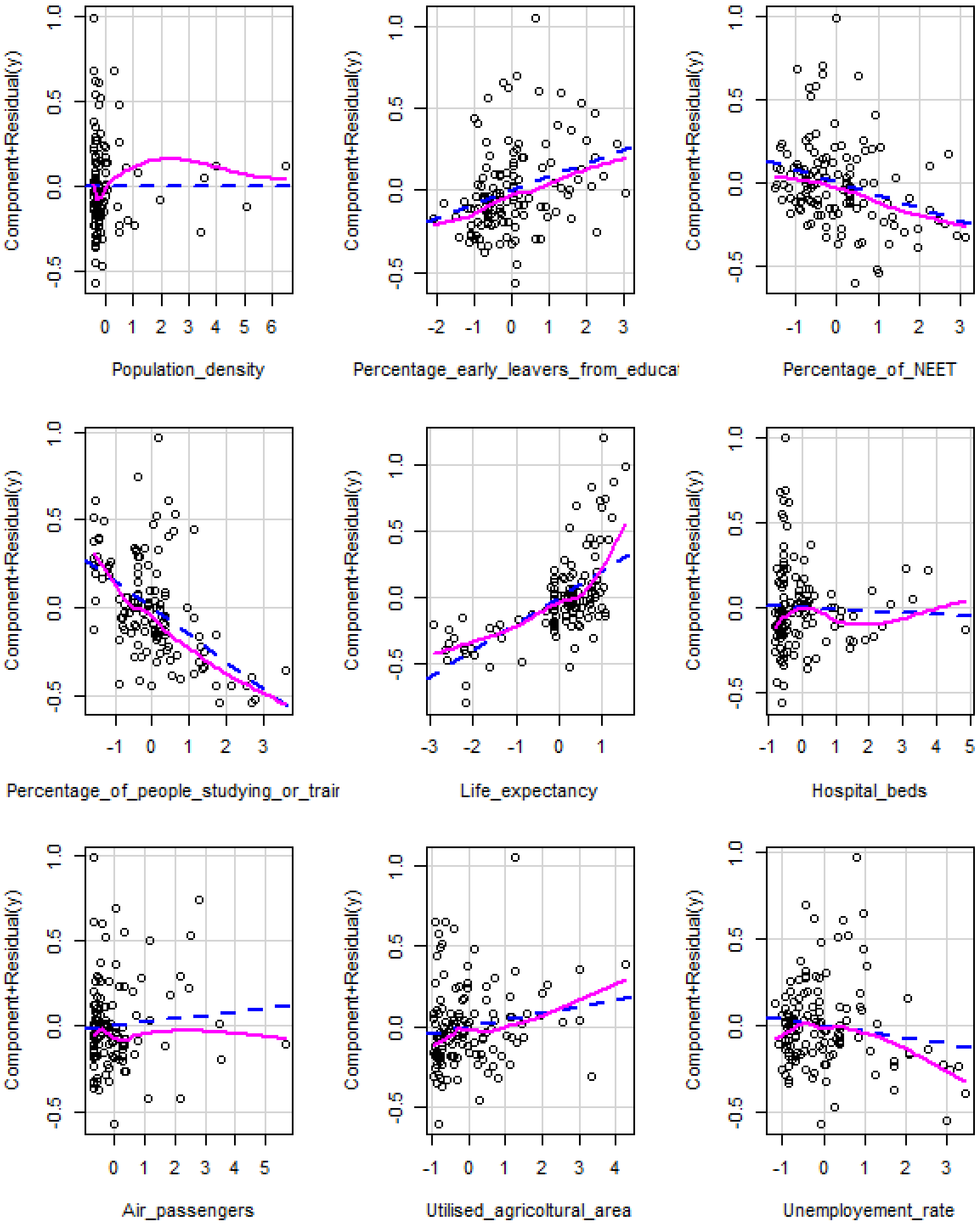


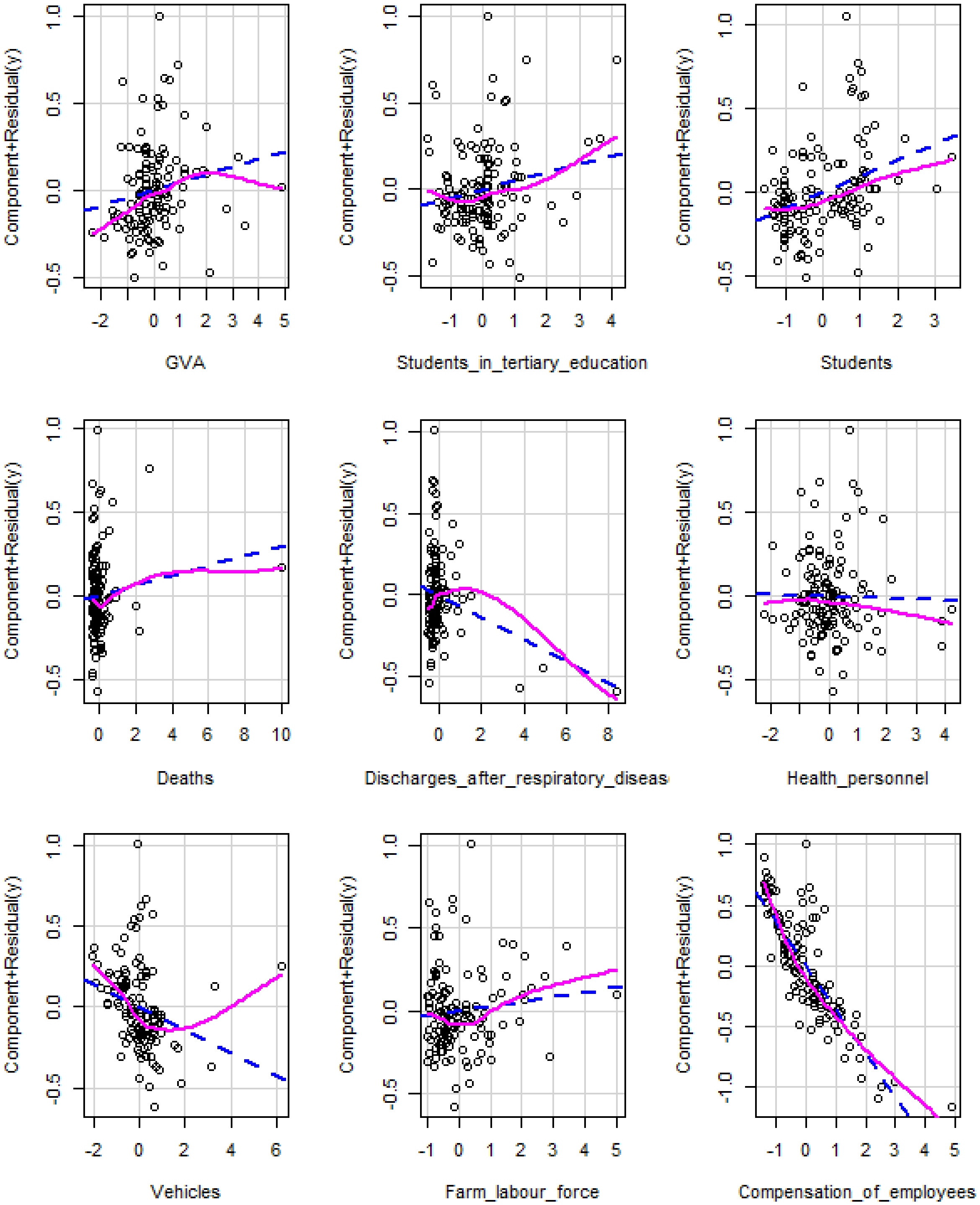
```
Call:
lm(formula = y ~ ., data = x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.56993 -0.15036 -0.03686  0.11086  0.99632 

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                0.3497437  0.0237777  14.709  < 2e-16 ***
Population_density                         0.0003239  0.0376538   0.009   0.9932    
Percentage_early_leavers_from_education    0.0840580  0.0354386   2.372   0.0195 *  
Percentage_of_NEET                        -0.0768999  0.0461211  -1.667   0.0984 .  
Percentage_of_people_studying_or_training -0.1518873  0.0488377  -3.110   0.0024 ** 
Life_expectancy                            0.1977369  0.0443411   4.459 2.04e-05 ***
Hospital_beds                             -0.0095925  0.0323706  -0.296   0.7675    
Air_passengers                             0.0205456  0.0266774   0.770   0.4429    
Utilised_agricoltural_area                 0.0412570  0.0335025   1.231   0.2209    
Unemployement_rate                        -0.0343063  0.0479144  -0.716   0.4756    
GVA                                        0.0447755  0.0306582   1.460   0.1471    
Students_in_tertiary_education             0.0490647  0.0326994   1.500   0.1364    
Students                                   0.0955859  0.0405970   2.355   0.0204 *  
Deaths                                     0.0301742  0.0298960   1.009   0.3151    
Discharges_after_respiratory_disease      -0.0672345  0.0292022  -2.302   0.0232 *  
Health_personnel                          -0.0054996  0.0316347  -0.174   0.8623    
Vehicles                                  -0.0706516  0.0413074  -1.710   0.0901 .  
Farm_labour_force                          0.0288683  0.0398469   0.724   0.4704    
Compensation_of_employees                 -0.3593071  0.1881834  -1.909   0.0589 .  
Hours_worked                               0.1162332  0.0474481   2.450   0.0159 *  
GDP                                        0.3639178  0.1910960   1.904   0.0595 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.269 on 107 degrees of freedom
Multiple R-squared:  0.4471,	Adjusted R-squared:  0.3437 
F-statistic: 4.326 on 20 and 107 DF,  p-value: 3.334e-07

Hit <Return> to see next plot: 
Hit <Return> to see next plot: 
Hit <Return> to see next plot: 

	Shapiro-Wilk normality test

data:  residuals(fm)
W = 0.9344, p-value = 9.981e-06

Linear Regression 

128 samples
 20 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation 
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ... 
Resampling results:

  RMSE       Rsquared   MAE      
  0.3505417  0.1070632  0.2350854
```

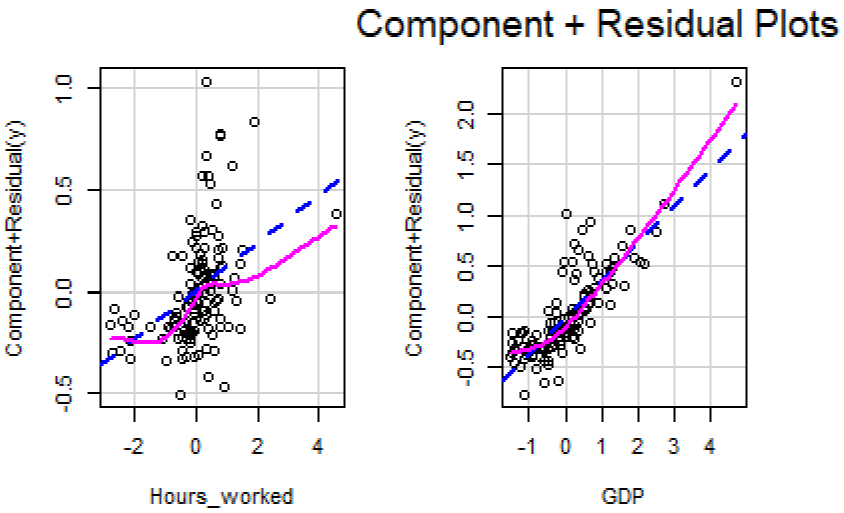




As we can see from Shapiro test there are statistical reasons to reject null hypothesis that residuals have normal distribution.

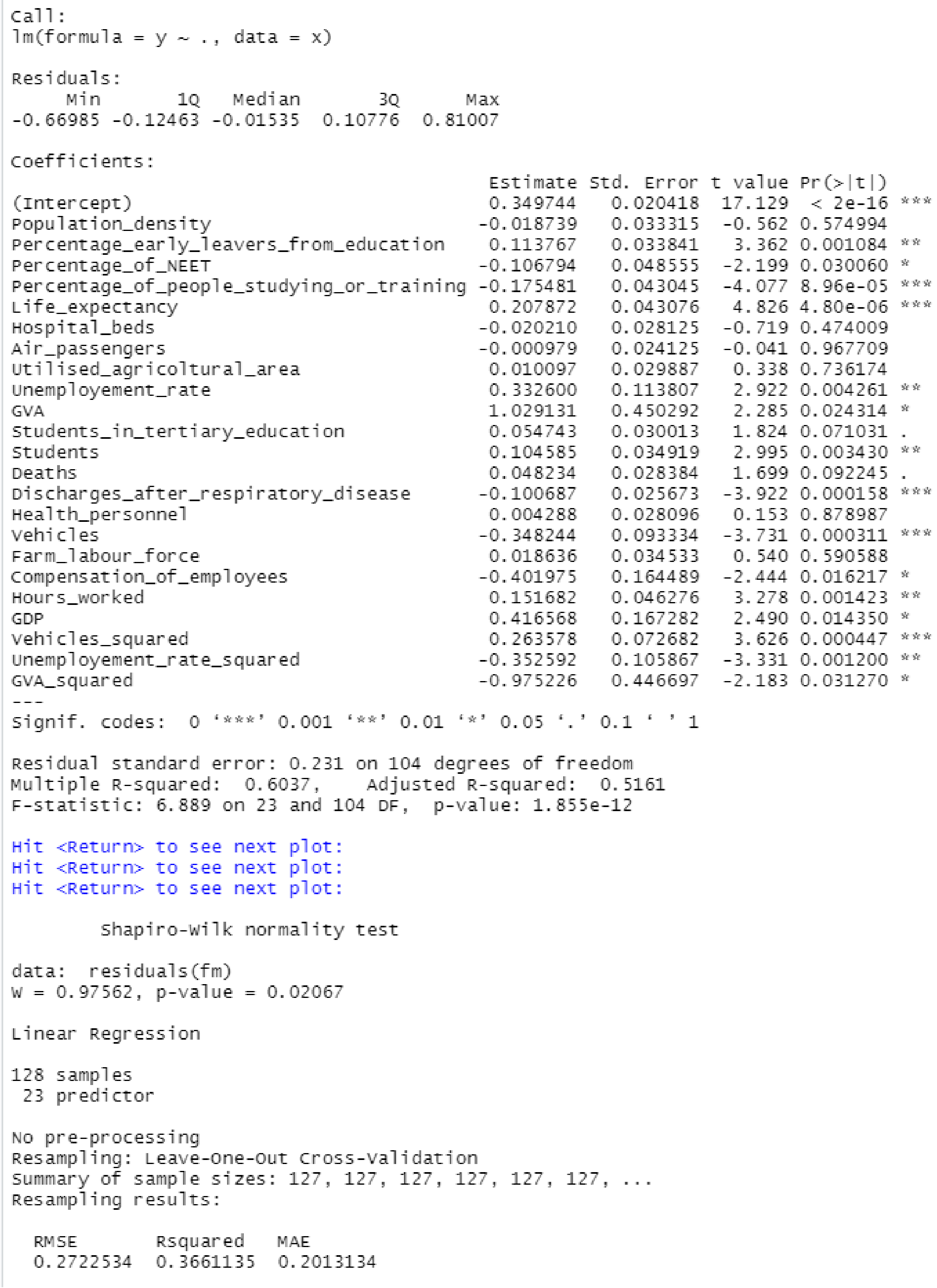

Component+Residual(y)
-0.5
0.0
0.5
1.0
0
1
2
3
4
5
6
Population_density
Component+Residual(y)
-0.5
0.0
0.5
1.0
-2
-1
0
1
2
3
Percentage_early_leavers_from_educa
Component+Residual(y)
-0.5
0.0
0.5
1.0
-1
0
1
2
3
Percentage_of_NEET
Component+Residual(y)
-0.5
0.0
0.5
1.0
-1
0
1
2
3
Percentage_of_people_studying_or_trair
Component+Residual(y)
-0.5
0.0
0.5
1.0
-3
-2
-1
0
1
Life_expectancy
Component+Residual(y)
-0.5
0.0
0.5
1.0
-1
0
1
2
3
4
5
Hospital_beds
Component+Residual(y)
-0.5
0.0
0.5
1.0
0
1
2
3
4
5
Air_passengers
Component+Residual(y)
-0.5
0.0
0.5
1.0
-1
0
1
2
3
4
Utilised_agricoltural_area
Component+Residual(y)
-0.5
0.0
0.5
1.0
-1
0
1
2
3
Unemployement_rate


Component+Residual(y)
1.0
0.5
0.0
-0.5
-2
0
1
2
3
4
5
GVA
Component+Residual(y)
1.0
0.5
0.0
-0.5
-1
0
1
2
3
4
Students_in_tertiary_education
Component+Residual(y)
1.0
0.5
0.0
-0.5
-1
0
1
2
3
Students
Component+Residual(y)
1.0
0.5
0.0
-0.5
0
2
4
6
8
10
Deaths
Component+Residual(y)
1.0
0.5
0.0
-0.5
0
2
4
6
8
Discharges_after_respiratory_diseas
Component+Residual(y)
1.0
0.5
0.0
-0.5
-2
0
1
2
3
4
Health_personnel
Component+Residual(y)
1.0
0.5
0.0
-0.5
-2
0
2
4
6
Vehicles
Component+Residual(y)
1.0
0.5
0.0
-0.5
-1
0
1
2
3
4
5
Farm_labour_force
Component+Residual(y)
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
2
3
4
5
Compensation_of_employees



Partial residual plots were used to detect a need for introduction of nonlinear features like squares of **Vehicles, Unemployement_rate, GVA.**

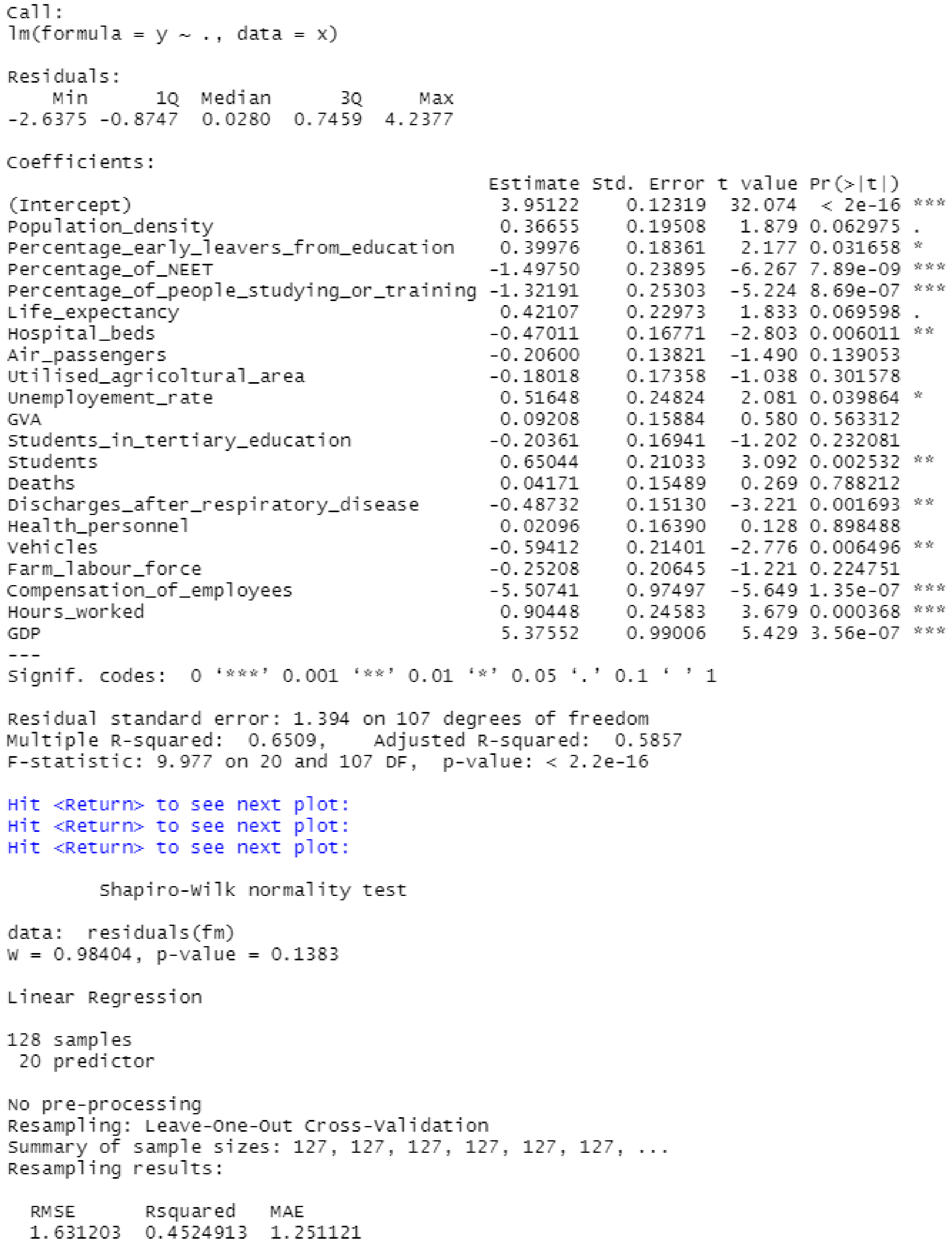
### Linear regression with all features + nonlinear features - First wave



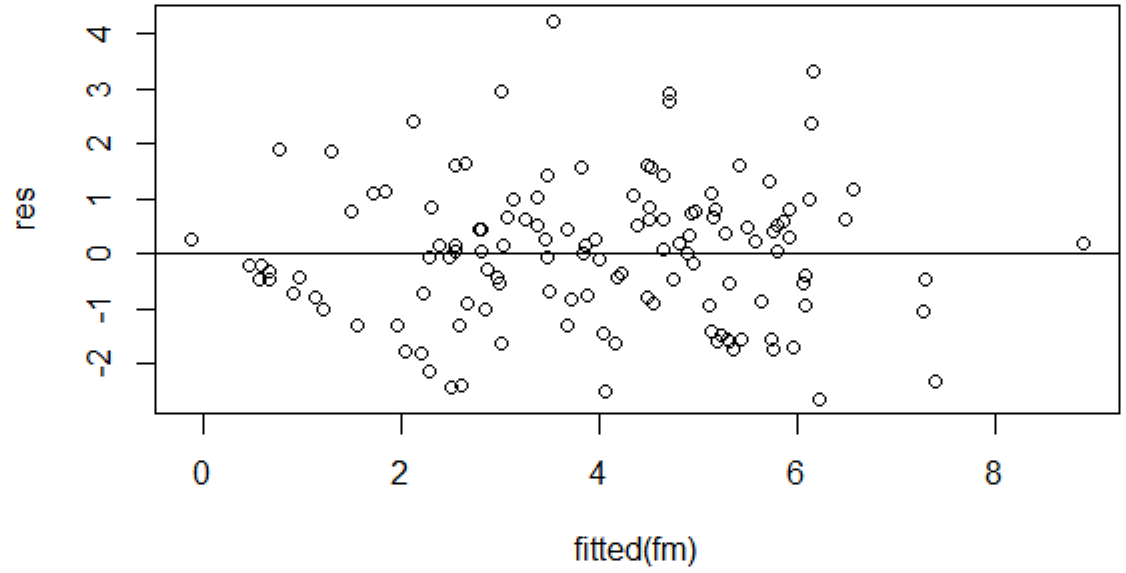
```
Call:
lm(formula = y ~ ., data = x)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.66985 -0.12463 -0.01535  0.10776  0.81007 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                               0.349744   0.020418  17.129  < 2e-16 ***
Population_density                       -0.018739   0.033315  -0.562 0.574994    
Percentage_early_leavers_from_education   0.113767   0.033841   3.362 0.001084 ** 
Percentage_of_NEET                       -0.106794   0.048555  -2.199 0.030060 *  
Percentage_of_people_studying_or_training -0.175481   0.043045  -4.077 8.96e-05 ***
Life_expectancy                           0.207872   0.043076   4.826 4.80e-06 ***
Hospital_beds                            -0.020210   0.028125  -0.719 0.474009    
Air_passengers                           -0.000979   0.024125  -0.041 0.967709    
Utilised_agricoltural_area                0.010097   0.029887   0.338 0.736174    
Unemployement_rate                        0.332600   0.113807   2.922 0.004261 ** 
GVA                                       1.029131   0.450292   2.285 0.024314 *  
Students_in_tertiary_education            0.054743   0.030013   1.824 0.071031 .  
Students                                  0.104585   0.034919   2.995 0.003430 ** 
Deaths                                    0.048234   0.028384   1.699 0.092245 .  
Discharges_after_respiratory_disease     -0.100687   0.025673  -3.922 0.000158 ***
Health_personnel                          0.004288   0.028096   0.153 0.878987    
Vehicles                                 -0.348244   0.093334  -3.731 0.000311 ***
Farm_labour_force                         0.018636   0.034533   0.540 0.590588    
Compensation_of_employees                -0.401975   0.164489  -2.444 0.016217 *  
Hours_worked                              0.151682   0.046276   3.278 0.001423 ** 
GDP                                       0.416568   0.167282   2.490 0.014350 *  
Vehicles_squared                          0.263578   0.072682   3.626 0.000447 ***
Unemployement_rate_squared               -0.352592   0.105867  -3.331 0.001200 ** 
GVA_squared                              -0.975226   0.446697  -2.183 0.031270 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.231 on 104 degrees of freedom
Multiple R-squared:  0.6037,	Adjusted R-squared:  0.5161 
F-statistic: 6.889 on 23 and 104 DF,  p-value: 1.855e-12

Hit <Return> to see next plot: 
Hit <Return> to see next plot: 
Hit <Return> to see next plot: 

	Shapiro-Wilk normality test

data:  residuals(fm)
W = 0.97562, p-value = 0.02067

Linear Regression 

128 samples
 23 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation 
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ... 
Resampling results:

  RMSE       Rsquared   MAE      
  0.2722534  0.3661135  0.2013134
```

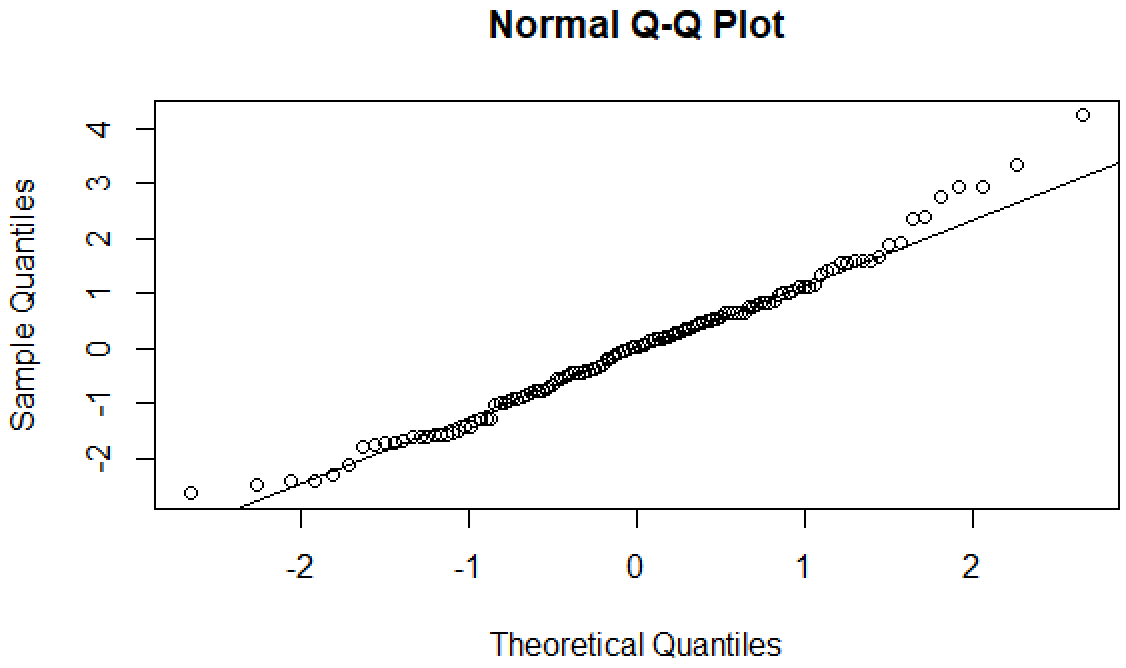
It can be seen that introduction of nonlinear features improved metrics estimated with LOOCV.

## Linear regression with all features - Second wave

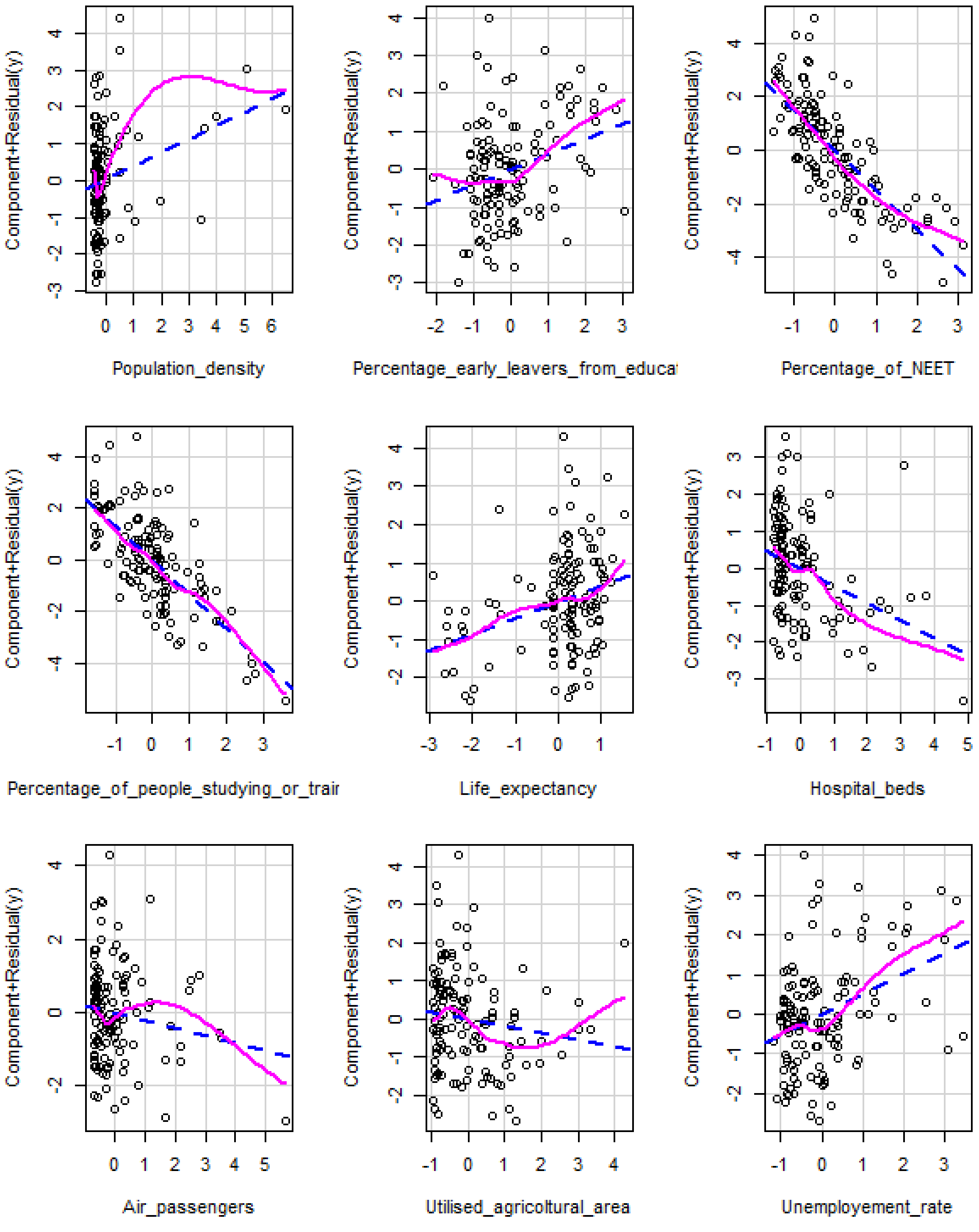


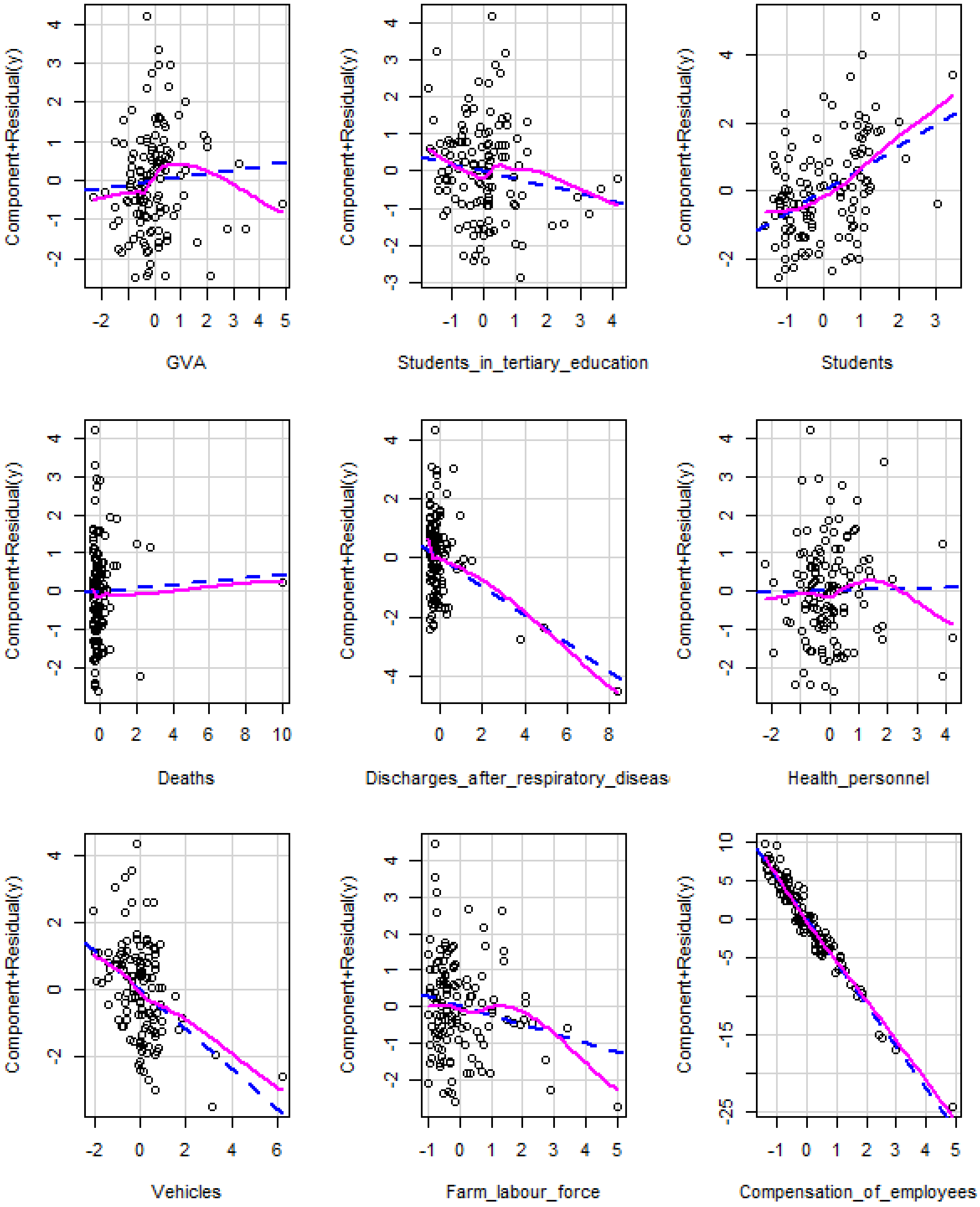
```
Call:
lm(formula = y ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6375 -0.8747  0.0280  0.7459  4.2377 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                3.95122    0.12319  32.074  < 2e-16 ***
Population_density                         0.36655    0.19508   1.879 0.062975 .  
Percentage_early_leavers_from_education    0.39976    0.18361   2.177 0.031658 *  
Percentage_of_NEET                        -1.49750    0.23895  -6.267 7.89e-09 ***
Percentage_of_people_studying_or_training -1.32191    0.25303  -5.224 8.69e-07 ***
Life_expectancy                            0.42107    0.22973   1.833 0.069598 .  
Hospital_beds                             -0.47011    0.16771  -2.803 0.006011 ** 
Air_passengers                            -0.20600    0.13821  -1.490 0.139053    
Utilised_agricoltural_area                -0.18018    0.17358  -1.038 0.301578    
Unemployement_rate                         0.51648    0.24824   2.081 0.039864 *  
GVA                                        0.09208    0.15884   0.580 0.563312    
Students_in_tertiary_education            -0.20361    0.16941  -1.202 0.232081    
Students                                   0.65044    0.21033   3.092 0.002532 ** 
Deaths                                     0.04171    0.15489   0.269 0.788212    
Discharges_after_respiratory_disease      -0.48732    0.15130  -3.221 0.001693 ** 
Health_personnel                           0.02096    0.16390   0.128 0.898488    
Vehicles                                  -0.59412    0.21401  -2.776 0.006496 ** 
Farm_labour_force                         -0.25208    0.20645  -1.221 0.224751    
Compensation_of_employees                 -5.50741    0.97497  -5.649 1.35e-07 ***
Hours_worked                               0.90448    0.24583   3.679 0.000368 ***
GDP                                        5.37552    0.99006   5.429 3.56e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.394 on 107 degrees of freedom
Multiple R-squared:  0.6509,	Adjusted R-squared:  0.5857 
F-statistic: 9.977 on 20 and 107 DF,  p-value: < 2.2e-16

Hit <Return> to see next plot: 
Hit <Return> to see next plot: 
Hit <Return> to see next plot: 

	Shapiro-Wilk normality test

data:  residuals(fm)
W = 0.98404, p-value = 0.1383

Linear Regression 

128 samples
 20 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation 
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ... 
Resampling results:

  RMSE      Rsquared   MAE     
  1.631203  0.4524913  1.251121
```

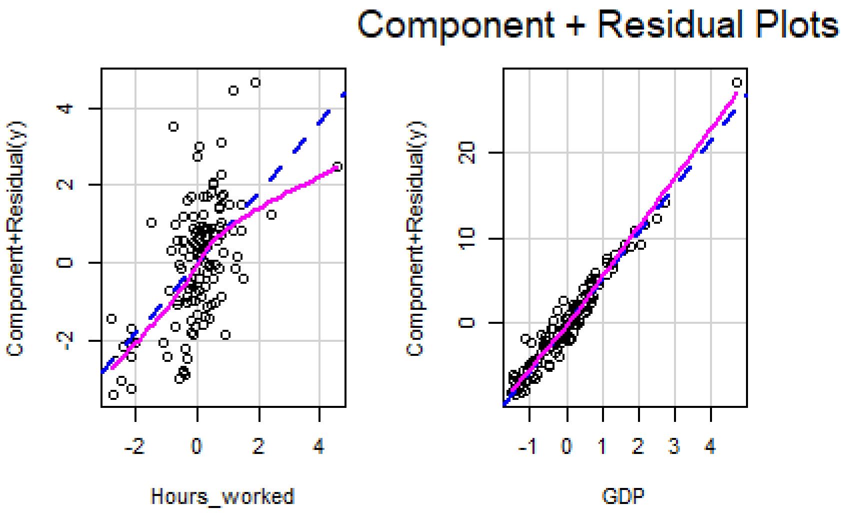




No statistical reason to reject the null hypothesis.

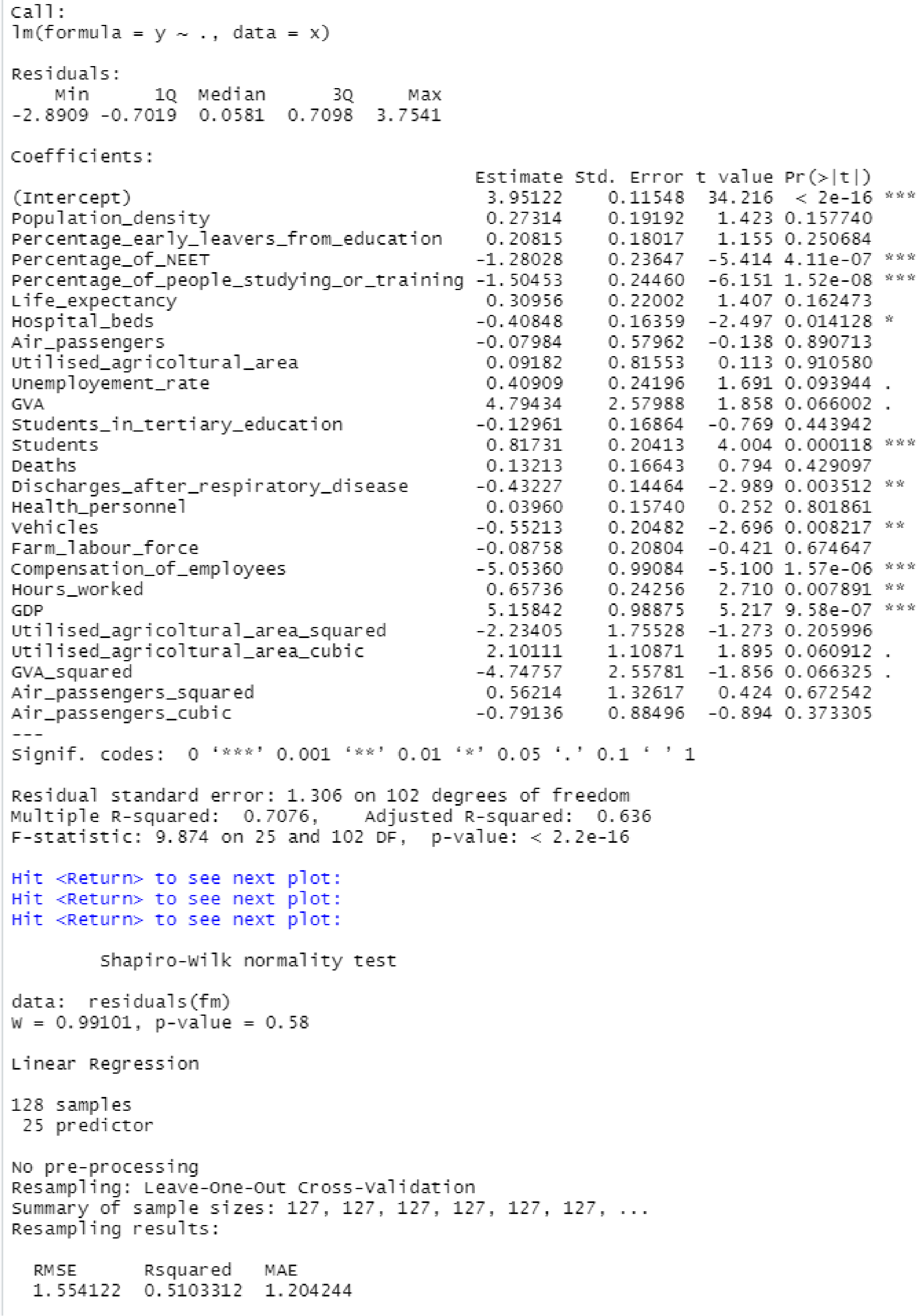

Component+Residual(y)
Population_density
0 1 2 3 4 5 6
-3 -2 -1 0 1 2 3 4
Percentage_early_leavers_from_educa
-2 -1 0 1 2 3
-3 -2 -1 0 1 2 3 4
Percentage_of_NEET
-1 0 1 2 3
-4 -2 0 2 4
Percentage_of_people_studying_or_trair
-1 0 1 2 3
-4 -2 0 2 4
Life_expectancy
-3 -2 -1 0 1
-2 -1 0 1 2 3 4
Hospital_beds
-1 0 1 2 3 4 5
-3 -2 -1 0 1 2 3
Air_passengers
0 1 2 3 4 5
-2 0 2 4
Utilised_agricoltural_area
-1 0 1 2 3 4
-2 -1 0 1 2 3 4
Unemployement_rate
-1 0 1 2 3
-2 -1 0 1 2 3 4


Component+Residual(y)
4
3
2
1
0
-1
-2
-2
0
1
2
3
4
5
GVA
Component+Residual(y)
4
3
2
1
0
-1
-2
-3
-1
0
1
2
3
4
Students_in_tertiary_education
Component+Residual(y)
4
2
0
-2
-1
0
1
2
3
Students
Component+Residual(y)
4
3
2
1
0
-1
-2
0
2
4
6
8
10
Deaths
Component+Residual(y)
4
2
0
-2
-4
0
2
4
6
8
Discharges_after_respiratory_diseas
Component+Residual(y)
4
3
2
1
0
-1
-2
-2
0
1
2
3
4
Health_personnel
Component+Residual(y)
4
2
0
-2
-2
0
2
4
6
Vehicles
Component+Residual(y)
4
3
2
1
0
-1
-2
-1
0
1
2
3
4
5
Farm_labour_force
Component+Residual(y)
10
5
0
-5
-15
-25
-1
0
1
2
3
4
5
Compensation_of_employees



Again partial residual plots were used to spot possible nonlinear features. Following features were introduced: **Utilised_agricoltural_area squared, Utilised_agricoltural_area cubed, GVA squared, Air_passengers squared, Air_passengers cubed.**

### Linear regression with all features + nonlinear features - Second wave

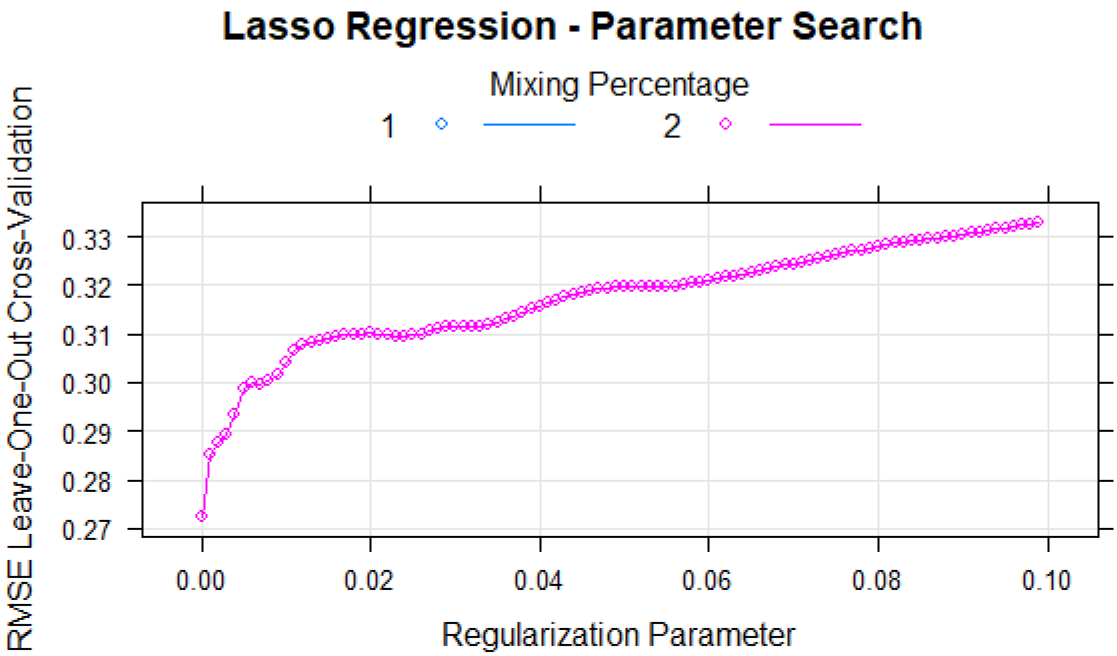


```
Call:
lm(formula = y ~ ., data = x)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8909 -0.7019  0.0581  0.7098  3.7541 

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                3.95122    0.11548  34.216  < 2e-16 ***
Population_density                         0.27314    0.19192   1.423 0.157740    
Percentage_early_leavers_from_education    0.20815    0.18017   1.155 0.250684    
Percentage_of_NEET                        -1.28028    0.23647  -5.414 4.11e-07 ***
Percentage_of_people_studying_or_training -1.50453    0.24460  -6.151 1.52e-08 ***
Life_expectancy                            0.30956    0.22002   1.407 0.162473    
Hospital_beds                             -0.40848    0.16359  -2.497 0.014128 *  
Air_passengers                            -0.07984    0.57962  -0.138 0.890713    
Utilised_agricoltural_area                 0.09182    0.81553   0.113 0.910580    
Unemployement_rate                         0.40909    0.24196   1.691 0.093944 .  
GVA                                        4.79434    2.57988   1.858 0.066002 .  
Students_in_tertiary_education            -0.12961    0.16864  -0.769 0.443942    
Students                                   0.81731    0.20413   4.004 0.000118 ***
Deaths                                     0.13213    0.16643   0.794 0.429097    
Discharges_after_respiratory_disease      -0.43227    0.14464  -2.989 0.003512 ** 
Health_personnel                           0.03960    0.15740   0.252 0.801861    
Vehicles                                  -0.55213    0.20482  -2.696 0.008217 ** 
Farm_labour_force                         -0.08758    0.20804  -0.421 0.674647    
Compensation_of_employees                 -5.05360    0.99084  -5.100 1.57e-06 ***
Hours_worked                               0.65736    0.24256   2.710 0.007891 ** 
GDP                                        5.15842    0.98875   5.217 9.58e-07 ***
Utilised_agricoltural_area_squared        -2.23405    1.75528  -1.273 0.205996    
Utilised_agricoltural_area_cubic           2.10111    1.10871   1.895 0.060912 .  
GVA_squared                               -4.74757    2.55781  -1.856 0.066325 .  
Air_passengers_squared                     0.56214    1.32617   0.424 0.672542    
Air_passengers_cubic                      -0.79136    0.88496  -0.894 0.373305    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.306 on 102 degrees of freedom
Multiple R-squared:  0.7076,	Adjusted R-squared:  0.636 
F-statistic: 9.874 on 25 and 102 DF,  p-value: < 2.2e-16

Hit <Return> to see next plot: 
Hit <Return> to see next plot: 
Hit <Return> to see next plot: 

	Shapiro-Wilk normality test

data:  residuals(fm)
W = 0.99101, p-value = 0.58

Linear Regression 

128 samples
 25 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation 
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ... 
Resampling results:

  RMSE      Rsquared   MAE     
  1.554122  0.5103312  1.204244
```

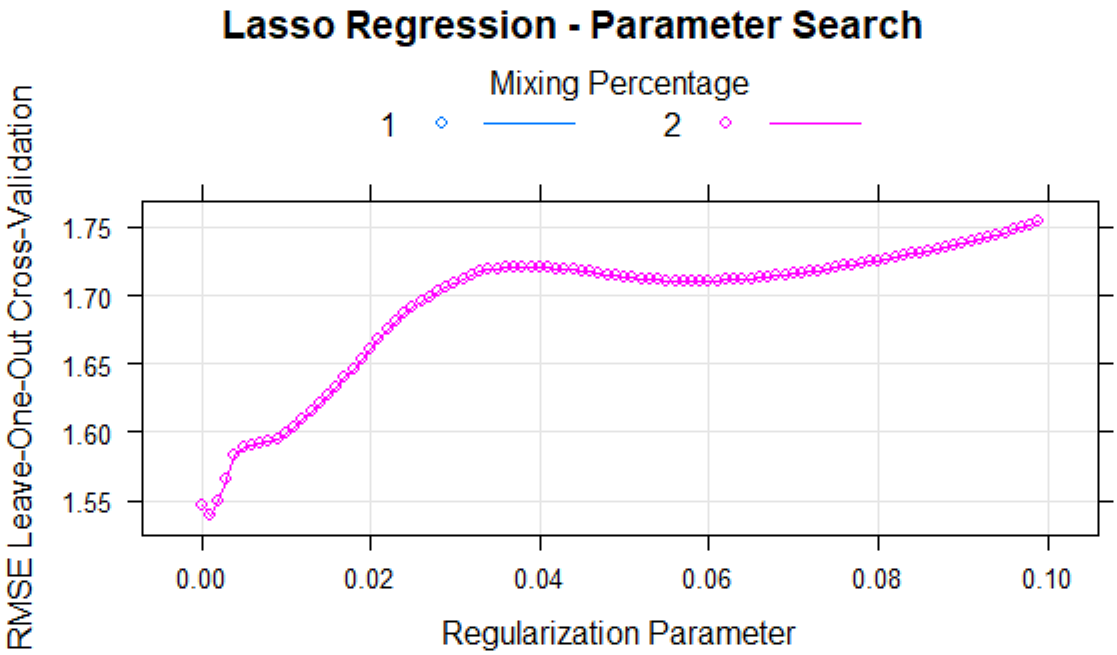
After introduction of nonlinear features again we can observe slightly improved estimated metrics.

### Lasso regression

**First wave**



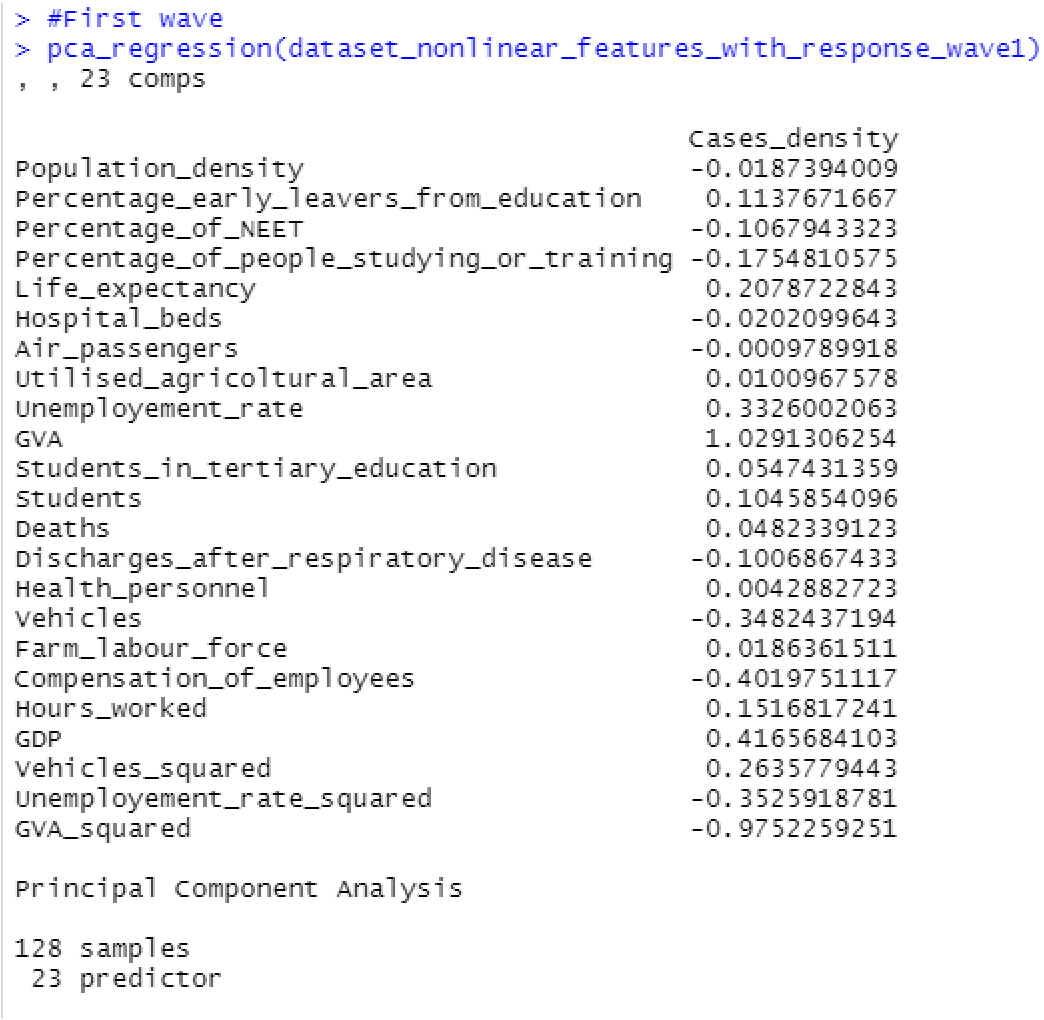
**Second wave**



After introduction of nonlinear features, addition of regularization only decreases performance of linear model, which might suggest that reduction of feature space is not necessary favourable.

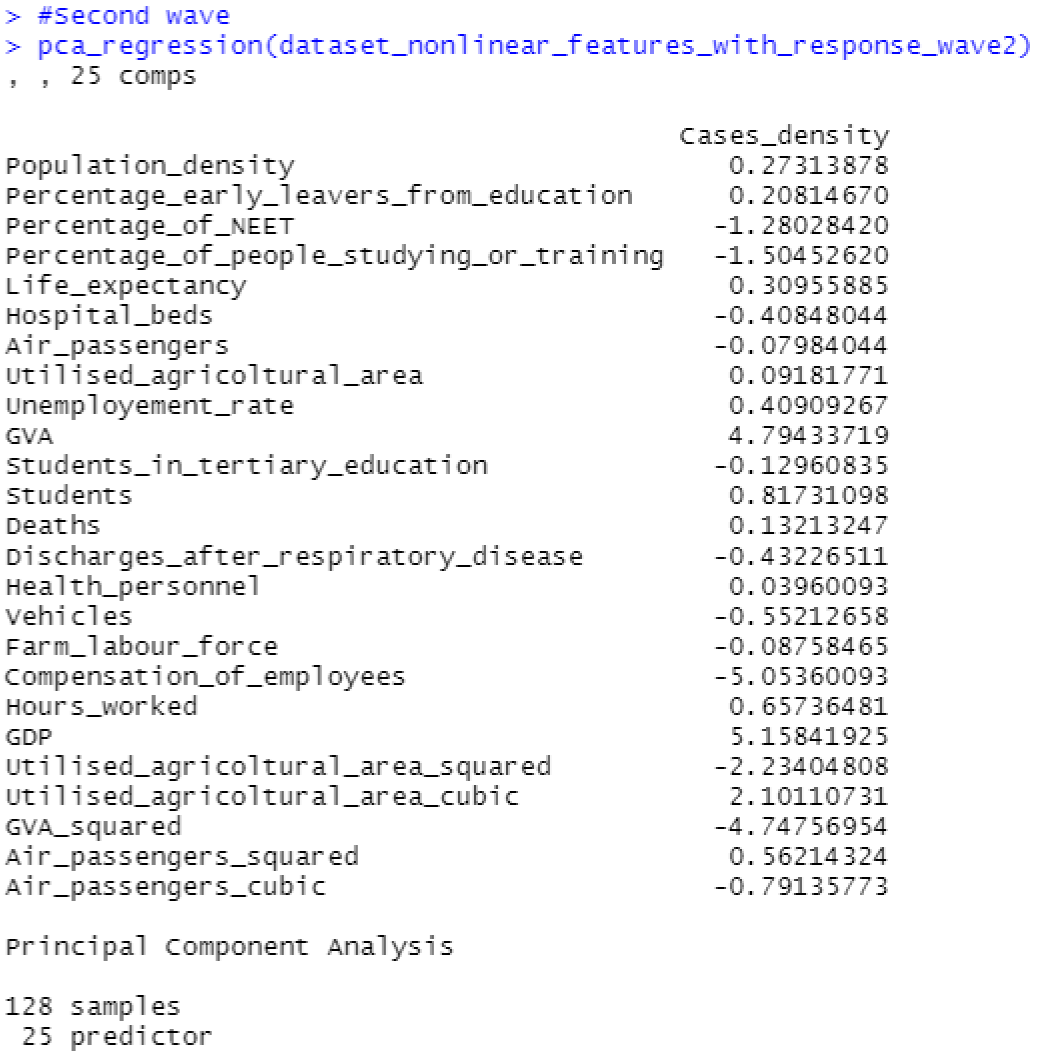
## PCA regression

### First wave



```
> #First wave
> pca_regression(dataset_nonlinear_features_with_response_wave1)
, , 23 comps

                                          Cases_density
Population_density                        -0.0187394009
Percentage_early_leavers_from_education    0.1137671667
Percentage_of_NEET                        -0.1067943323
Percentage_of_people_studying_or_training -0.1754810575
Life_expectancy                            0.2078722843
Hospital_beds                             -0.0202099643
Air_passengers                            -0.0009789918
Utilised_agricoltural_area                 0.0100967578
Unemployement_rate                         0.3326002063
GVA                                        1.0291306254
Students_in_tertiary_education             0.0547431359
Students                                   0.1045854096
Deaths                                     0.0482339123
Discharges_after_respiratory_disease      -0.1006867433
Health_personnel                           0.0042882723
Vehicles                                  -0.3482437194
Farm_labour_force                          0.0186361511
Compensation_of_employees                 -0.4019751117
Hours_worked                               0.1516817241
GDP                                        0.4165684103
Vehicles_squared                           0.2635779443
Unemployement_rate_squared                -0.3525918781
GVA_squared                               -0.9752259251

Principal Component Analysis

128 samples
 23 predictor
```

**Second wave**

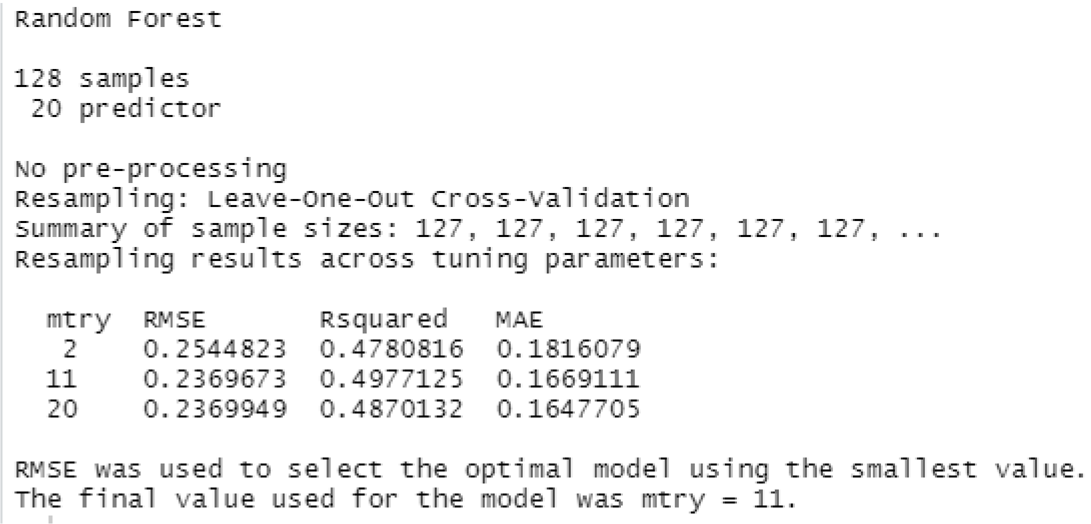


```
> #Second wave
> pca_regression(dataset_nonlinear_features_with_response_wave2)
, , 25 comps

                                          Cases_density
Population_density                          0.27313878
Percentage_early_leavers_from_education     0.20814670
Percentage_of_NEET                         -1.28028420
Percentage_of_people_studying_or_training  -1.50452620
Life_expectancy                             0.30955885
Hospital_beds                              -0.40848044
Air_passengers                             -0.07984044
Utilised_agricoltural_area                  0.09181771
Unemployement_rate                          0.40909267
GVA                                         4.79433719
Students_in_tertiary_education             -0.12960835
Students                                    0.81731098
Deaths                                      0.13213247
Discharges_after_respiratory_disease       -0.43226511
Health_personnel                            0.03960093
Vehicles                                   -0.55212658
Farm_labour_force                          -0.08758465
Compensation_of_employees                  -5.05360093
Hours_worked                                0.65736481
GDP                                         5.15841925
Utilised_agricoltural_area_squared         -2.23404808
Utilised_agricoltural_area_cubic            2.10110731
GVA_squared                                -4.74756954
Air_passengers_squared                      0.56214324
Air_passengers_cubic                       -0.79135773

Principal Component Analysis

128 samples
 25 predictor
```

Usage of Principal Components Regression is advantageous comparing to the usual linear regression, because it allows for straightforward interpretation of coefficients thanks to the process of making training dataset orthogonal, while in linear regression coefficients might be dependent on each other making interpretation of coefficients unpractical.

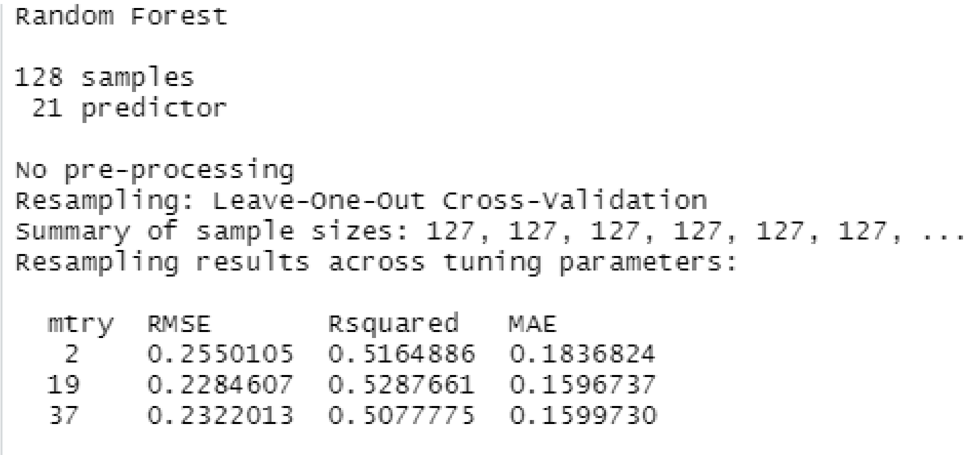
## Random forest

### First wave



```
Random Forest

128 samples
 20 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ...
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
   2    0.2544823  0.4780816  0.1816079
  11    0.2369673  0.4977125  0.1669111
  20    0.2369949  0.4870132  0.1647705

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was mtry = 11.
```

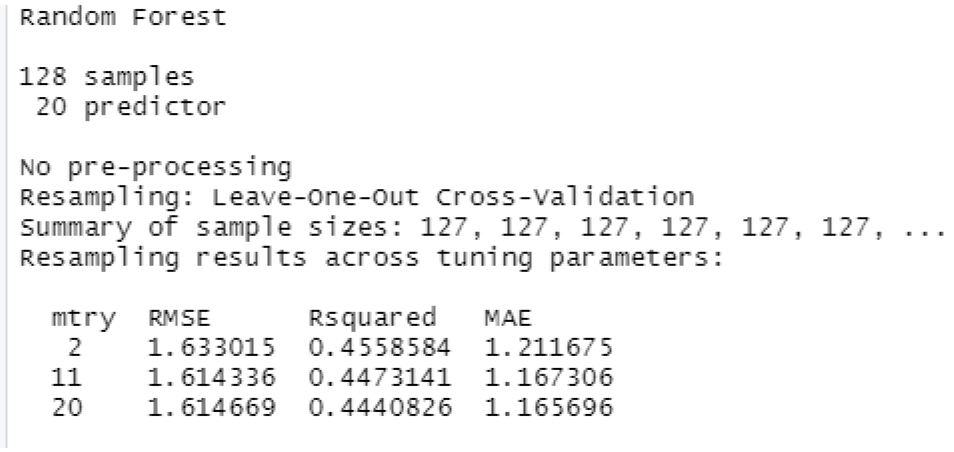
### First wave with country included as a feature



```
Random Forest

128 samples
 21 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ...
Resampling results across tuning parameters:

  mtry  RMSE       Rsquared   MAE
   2    0.2550105  0.5164886  0.1836824
  19    0.2284607  0.5287661  0.1596737
  37    0.2322013  0.5077775  0.1599730
```

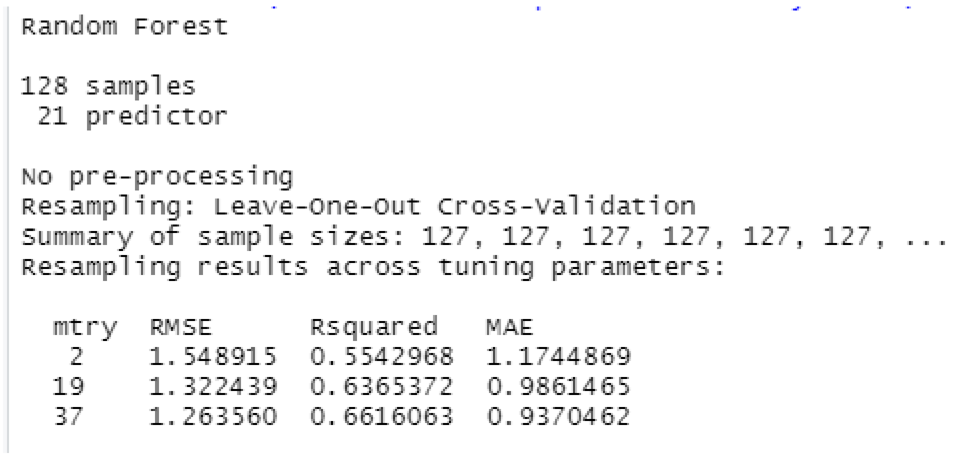
Inclusion of country might have improved estimated performance metrics just slightly, therefore possibly it is not such a relevant factor as it was thought at the beginning.

### Second wave



```
Random Forest

128 samples
 20 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    1.633015  0.4558584  1.211675
  11    1.614336  0.4473141  1.167306
  20    1.614669  0.4440826  1.165696
```

### Second wave with country included as a feature



```
Random Forest

128 samples
 21 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 127, 127, 127, 127, 127, 127, ...
Resampling results across tuning parameters:

  mtry  RMSE      Rsquared   MAE
   2    1.548915  0.5542968  1.1744869
  19    1.322439  0.6365372  0.9861465
  37    1.263560  0.6616063  0.9370462
```

In case of a second wave, however, inclusion of a country significantly reduced error, which is quite logical as after hit of the first wave countries started to introduce policies against spread of virus.

## Gradient Boosting

Gradient boosting methods are usually very good methods for tabular data, cause they reduce bias while remaining relatively low variance. Search of hyperparameters wasn’t very exhaustive so probably there is still a room for improvement.

## Comparison of regression methods

| **Method** | **MAE – 1st wave** | **MAE – 2nd wave** |
|---|---|---|
| **Linear regression** | 0.235 | 1.251 |
| **Linear regression + nonlinear features** | 0.201 | 1.204 |
| **Random forest** | 0.16 | 1.165 |
| **Random forest + country included** | 0.164 | **0.937** |
| **Gradient Boosting** | **0.147** | 1.255 |
| **Gradient Boosting + country included** | 0.156 | 1.263 |