Vilniaus Universitetas

Netiesinė regresija

Laboratorinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2022

Turinys

Naudoti metodai

Šiame darbe naudotas apibendrintas adityvus modelis su glodniaisiais splainais. Darbas atliktas naudojant R.

Naudoti R paketai:

tidyverse

rsample

corrplot

car

mgcv

gratia

effect

Duomenys ir jų šaltiniai

Betono mišinio stiprio duomenys pagal betono mišinį sudarančias medžiagas.

Duomenų šaltinis – UCI Machine Learning Repository. Prieiga per internetą:

https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength.

„cement“ - Cemento kiekis (kg viename m^3 mišinio).

„blast_furnace_slag“- aukštakrosnių šlako (kg viename m^3 mišinio).

„fly_ash“ - Lakiųjų pelenų kiekis (kg viename m^3 mišinio).

„water“ - Vandens kiekis (kg viename m^3 mišinio).

“superplasticizer” - Superplastiklių kiekis (kg viename m^3 mišinio).

“coarse_aggregate” - stambiojo užpildo kiekis betono mišinyje (kg viename m^3 mišinio).

„fine_aggregate“ - smulkiojo (smėlio) užpildo kiekis betono mišinyje (kg viename m^3 mišinio).

„age“ - Amžius (dienomis).

„strength“ - Cemento stipris gniaužiant (viekiant kompresinė spauda) (MPa) (atsako kintamasis).

Tikslas ir uždaviniai

Tikslas: Sudaryti netiesinį regresijos modelį betono stipriui prognozuoti pagal jo sudėties mišinį, kuris reikšmes prognozuotų tiksliau už tiesinį regresijos modelį.

Uždaviniai:

Tiesinio regresijos modelio betono stipriui sudarymas.

Tinkamų netiesinės regresijos metodų pasirinkimas turimai duomenų aibei.

Netiesinio regresijos modelio sudarymas.

Modelio tinkamumo analizė.

Tiesinio ir netiesinio modelių palyginimas.

Modelio panaudojimas prognozuoti betono stiprį esant tam tikrai mišinio sudėčiai.

Atliktos analizės aprašymas

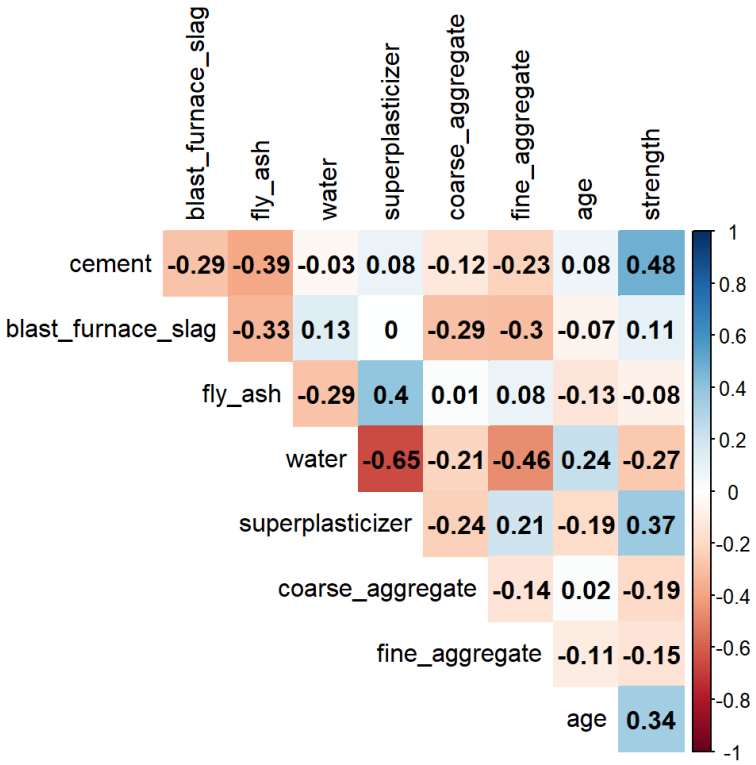
1. Naudojant R

Duomenų aibę apskritai sudaro 1030 stebėjų, tačiau 25 iš jų yra pasikartojantys (pagal visus požymius). Laikyta, kad šitie pasikartojantys stebėjimai neturėtų įtakoti modeliuojamo sąryšio tarp kovariančių ir atsako kintamojo, todėl pasirinkta šiuos stebėjimus pašalinti. Duomenų aibėje nėra praleistų reikšmių. Duomenys padalinti į mokymo ir testavimo aibes naudojant 75-25 santykį.

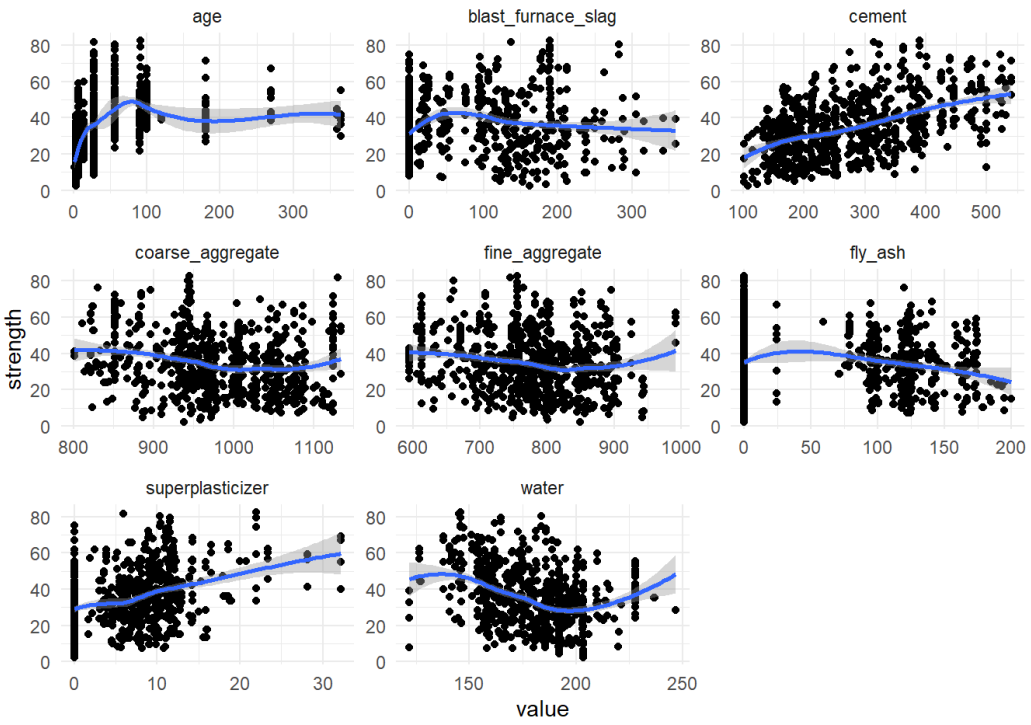
Nagrinėjant Pirsono koreliacijas tarp betono stiprio ir kovariančių vidutinė arba stipri koreliacija rasta tik su maža dalimi požymių. Kadangi Pirsono koreliacija matuoja tik tiesinį sąryšį tarp kintamųjų, iš prieš tai gauto rezultato pasirinkta laikyti, kad betono stipris gali būti ir netiesinė savo mišinio sudedamųjų dalių funkcija.

Panašus rezultatas gautas ir kiekvienai kovariantei ir atsakui nubraižius sklaidos diagramas kartu su netiesinės regresijos kreive. Šie gauti rezultatai neatsižvelgia į kitų kovariančių reikšmes, todėl jie gali tik sufleruoti apie galimą sąryšį tarp kovariantės ir atsako pilname regresijos modelyje (kuriame atsižvelgiama į kitų kovariančių reikšmes).

```
# Duomenų šaltinis
# https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength
# (https://www.kaggle.com/datasets/elikplim/concrete-compressive-strength-data-set)

library(tidyverse)
library(rsample)

concrete <- read_csv("concrete_data.csv")

concrete <- concrete %>%
  unique() %>%
  rename(strength = concrete_compressive_strength)

# Padalijama į mokymo ir testavimo aibes
set.seed(123)
concrete_split <- initial_split(concrete)
concrete_train <- training(concrete_split)
concrete_test <- testing(concrete_split)

library(corrplot)

correlation_matrix <- concrete_train %>%
  cor()

corrplot(correlation_matrix, order = "original", method = "color", type = "upper", diag =
FALSE, tl.col = "black", addCoef.col = "black")
```

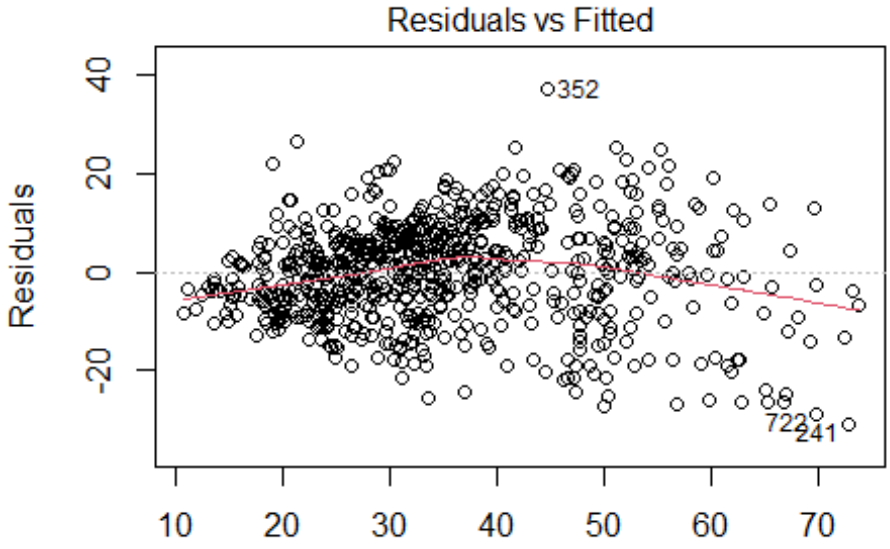



```
# tarp kovariančių ir atsako matome galimus netiesinis sąryšius
concrete_train %>%
  pivot_longer(-strength) %>%
  ggplot(aes(value, strength)) +
  geom_point() +
  facet_wrap(vars(name), scales = "free") +
  geom_smooth() +
  theme_minimal()
```

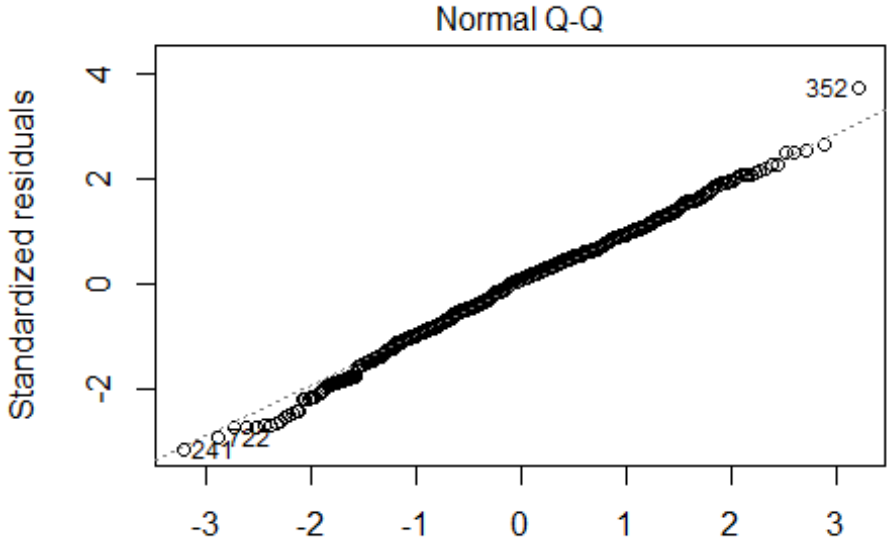


Pirmiausiai sudarytas paprastas tiesinės regresijos modelis. Analizuojant modelio diagnostinius grafikus rastas pakankamai stiprus liekanų heteroskedatiškumas. Pagal dalinių liekanų grafikus (partial residual plot / component residual plot) rasta, kad požymiuose “water” “superplastisizer” “age” turima papildoma informacija, į kurią paprastas tiesinis modelis nesugebėjo atsižvelgti.

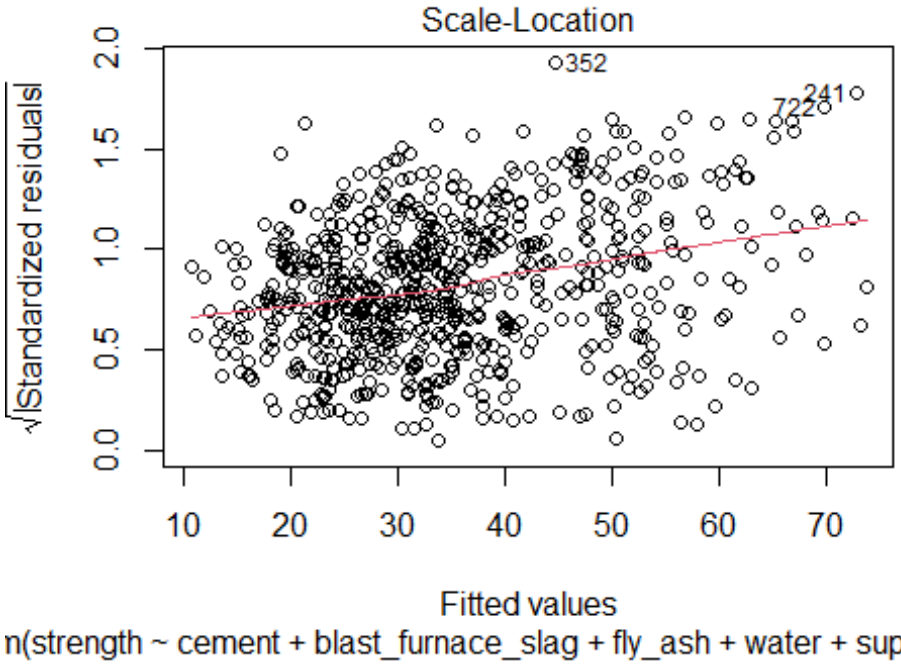
```
# Pirma sudaromas regresijos modelis su vien tiesiniais
  # kovariančių ir atsako sąryšiais (lyginamasis)
library(car)

baseline <- lm(strength ~ cement + blast_furnace_slag + fly_ash +
  water + superplasticizer
  + coarse_aggregate + fine_aggregate + age, data = concrete_train)


plot(baseline)
```

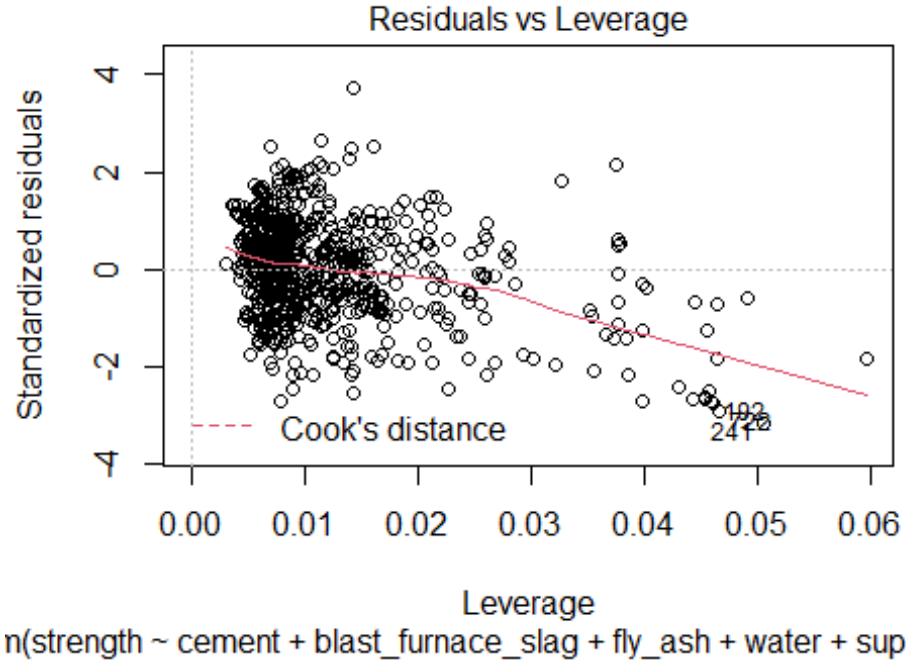


n(strength ~ cement + blast_furnace_slag + fly_ash + water + superpla

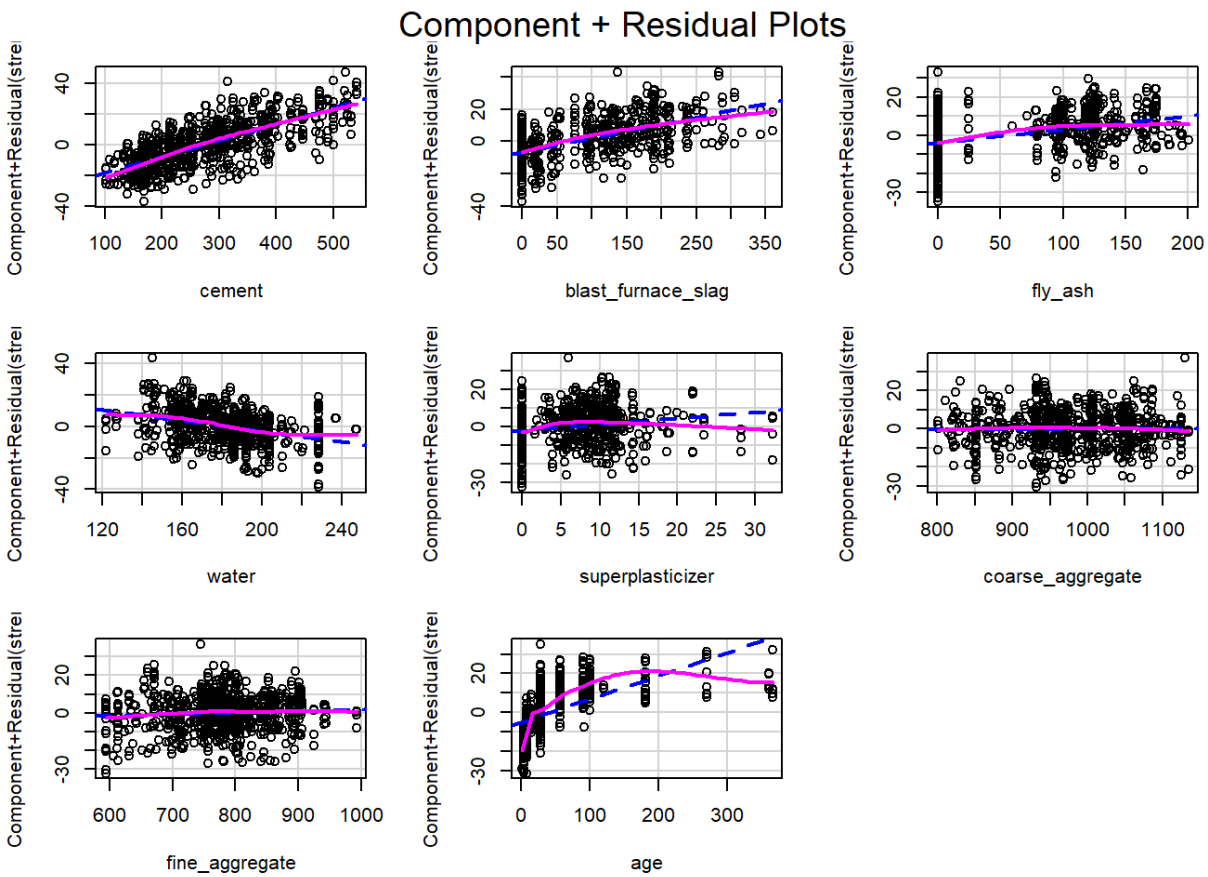


n(strength ~ cement + blast_furnace_slag + fly_ash + water + superpla



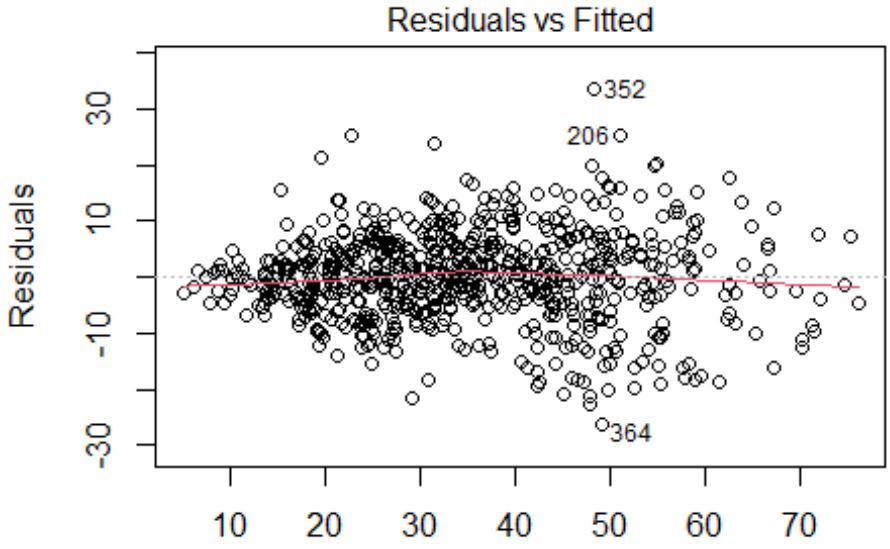


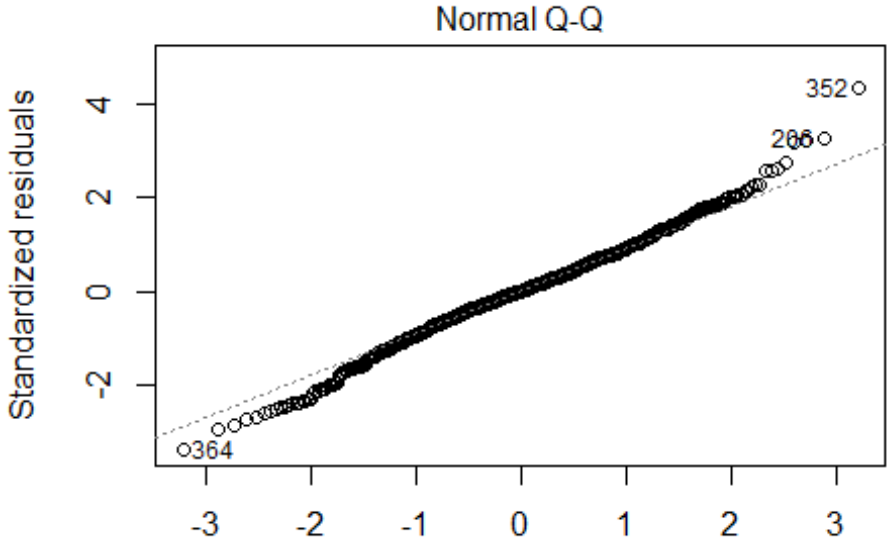

```
crPlots(baseline)
```

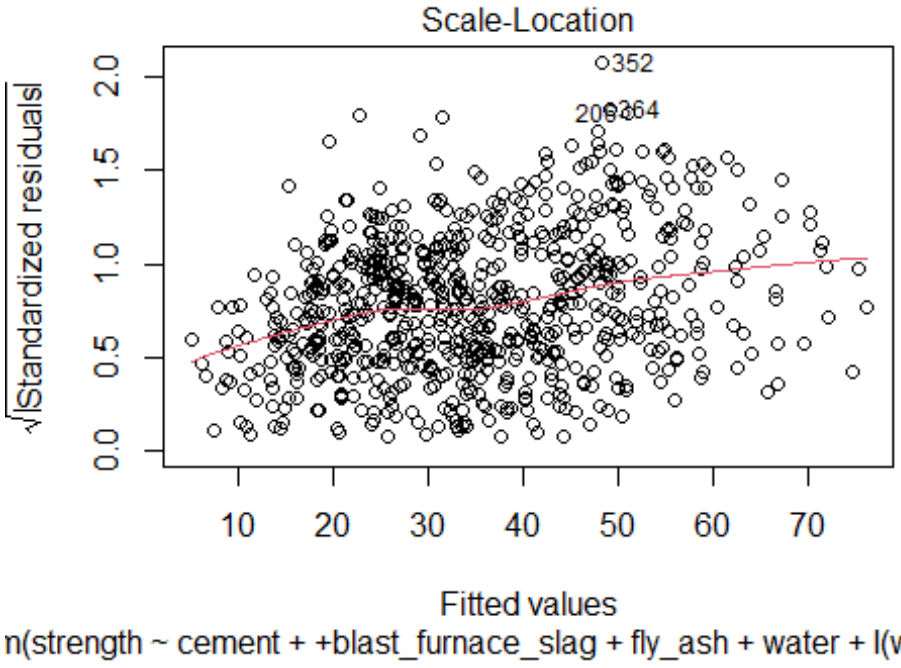


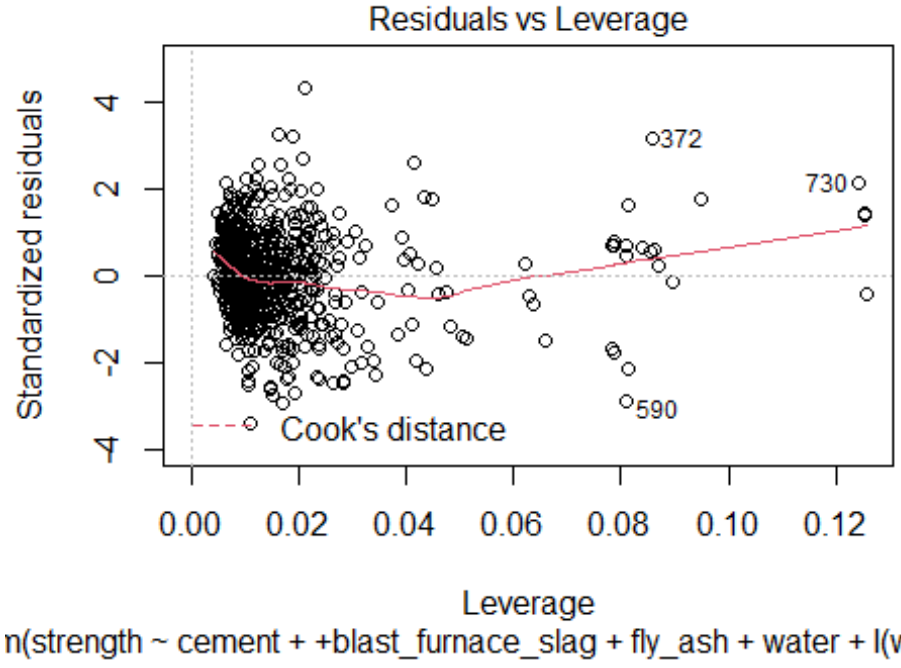
Siekiant patobulinti gautus rezultatus, sudarytas kitas regresijos modelis, papildomai į modelį įtraukiant prieš tai paminėtų kovariančių antruosius laipsnius (polinominės regresijos modelis). Naudojantis tomis pačiomis diagnostinėmis priemonėmis, laikyta, kad prieš tai rasti modelio nukrypimai nuo teorinio sumažėjo arba buvo panaikinti. Naudojant hierarchinių modelių (nested models) palyginimo testą, gautas statistiškai reikšmingas antro modelio skirtumas nuo sudaryto prieš tai ($p<0.001$).

```
# "naivus" metodas pagerinti modelį pridedant aukštesnius kovariančių laipsnius
model_polynomial <- lm(strength ~ cement + blast_furnace_slag
  + fly_ash + water + I(water^2)
  + superplasticizer + I(superplasticizer^2)
  + coarse_aggregate + fine_aggregate + age + I(age^2),
data = concrete_train
)

plot(model_polynomial)
```


Residuals vs Fitted
Residuals
Fitted values
n(strength ~ cement + +blast_furnace_slag + fly_ash + water + I(water′
352
206
364
30
10
-10
-30
10
20
30
40
50
60
70

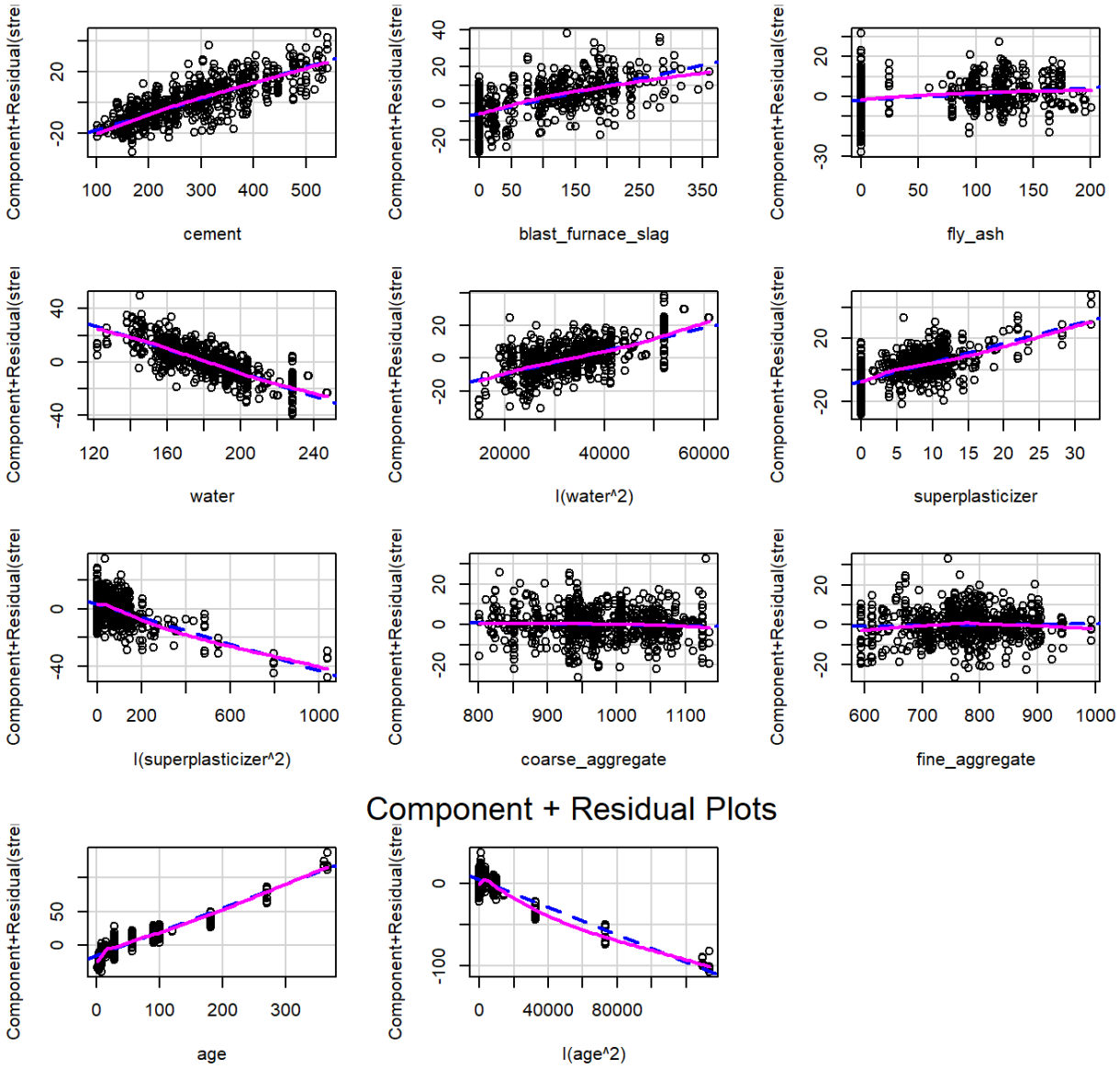

Normal Q-Q
Standardized residuals
Theoretical Quantiles
n(strength ~ cement + +blast_furnace_slag + fly_ash + water + I(water′
352
206
364
4
2
0
-2
-3
-2
-1
0
1
2
3






```
summary(model_polynomial)

##
## Call:
## lm(formula = strength ~ cement + +blast_furnace_slag + fly_ash +
```

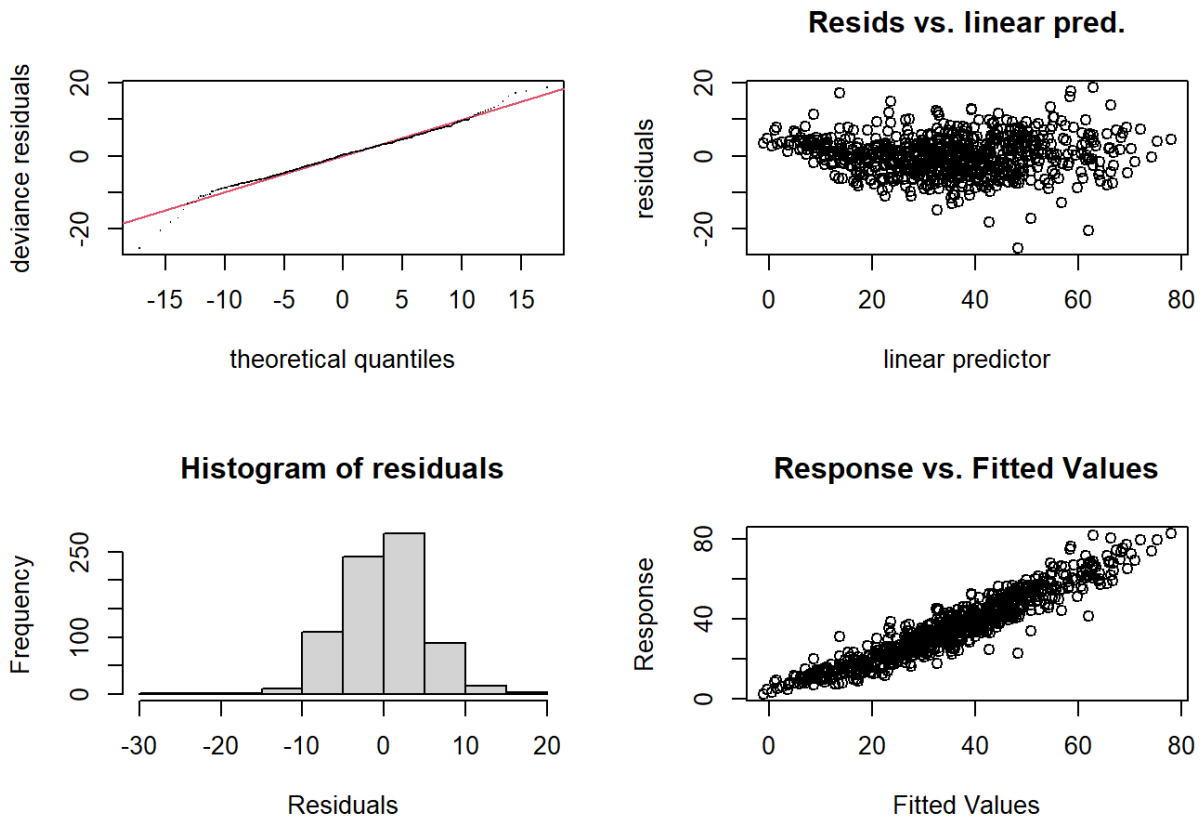
```
##     water + I(water^2) + superplasticizer + I(superplasticizer^2) +
##     coarse_aggregate + fine_aggregate + age + I(age^2), data = concrete_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.172  -4.597   0.047   4.843  33.480
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            4.325e+01  2.519e+01   1.717  0.08639 .
## cement                 1.025e-01  7.386e-03  13.870  < 2e-16 ***
## blast_furnace_slag     7.631e-02  8.966e-03   8.511  < 2e-16 ***
## fly_ash                3.166e-02  1.147e-02   2.761  0.00591 **
## water                 -4.414e-01  1.716e-01  -2.571  0.01032 *
## I(water^2)             6.992e-04  4.763e-04   1.468  0.14259
## superplasticizer       1.209e+00  1.494e-01   8.095 2.35e-15 ***
## I(superplasticizer^2) -4.655e-02  5.935e-03  -7.843 1.53e-14 ***
## coarse_aggregate      -5.052e-03  8.459e-03  -0.597  0.55051
## fine_aggregate         3.445e-03  9.486e-03   0.363  0.71660
## age                    3.527e-01  1.255e-02  28.095  < 2e-16 ***
## I(age^2)              -8.310e-04  4.081e-05 -20.364  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.825 on 741 degrees of freedom
## Multiple R-squared:  0.7635, Adjusted R-squared:   0.76
## F-statistic: 217.5 on 11 and 741 DF,  p-value: < 2.2e-16

crPlots(model_polynomial)
```




```
# modelis reikšmingai skiriasi nuo modelio be aukštesnio laipsnio narių
anova(baseline, model_polynomial)

## Analysis of Variance Table
##
## Model 1: strength ~ cement + blast_furnace_slag + fly_ash + water + superplasticizer +
##     coarse_aggregate + fine_aggregate + age
## Model 2: strength ~ cement + +blast_furnace_slag + fly_ash + water + I(water^2) +
##     superplasticizer + I(superplasticizer^2) + coarse_aggregate +
##     fine_aggregate + age + I(age^2)
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    744 75286
## 2    741 45374  3     29912 162.83 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Siekiant dar tiksliau modeliuoti betono stiprį pasirinkta sudaryti apibendrintą adityvų regresijos modelį, naudojant glodniaisius splainus. Šiam tikslui pasitelkta *mgcv* biblioteka. Naudojant *gam* funkciją ir jios bibliotekos parametras λ parenakamas auitomatiškai, naudojant generalized cross validation. Imant numatytąjį mazgų skaičių kiekvienai kovariantei įprastuose diagnostinuose grafikuose stiprių nukrypimų nerasta, tačiau tiek statistiniais testais, tiek naudojantis modelio kovariančių efektų grafikais pastebėta, kad pradinis mazgų skaičius gali būti per mažas tinkamai įvertinti sąryšius tarp kovariančių ir atsako (stipriausiai tai matoma su kovariante „coarse_aggregate“)

```
c: 29.5 on 10 and 742 DF,  p-value: < 2.2e-16

library(mgcv)
library(gratia)
# Alternatyvus modelis: apibendrintas adityvus modelis su glodniaisiais splainais

# lambda parametras parenkamas automatiškai pagal generalized cross-validation
model_gam <- gam(strength ~ s(cement) + s(blast_furnace_slag) + s(fly_ash)
  + s(water) + s(superplasticizer)
  + s(coarse_aggregate) + s(fine_aggregate) + s(age),
data = concrete_train,
select = TRUE
)

gam.check(model_gam)
```



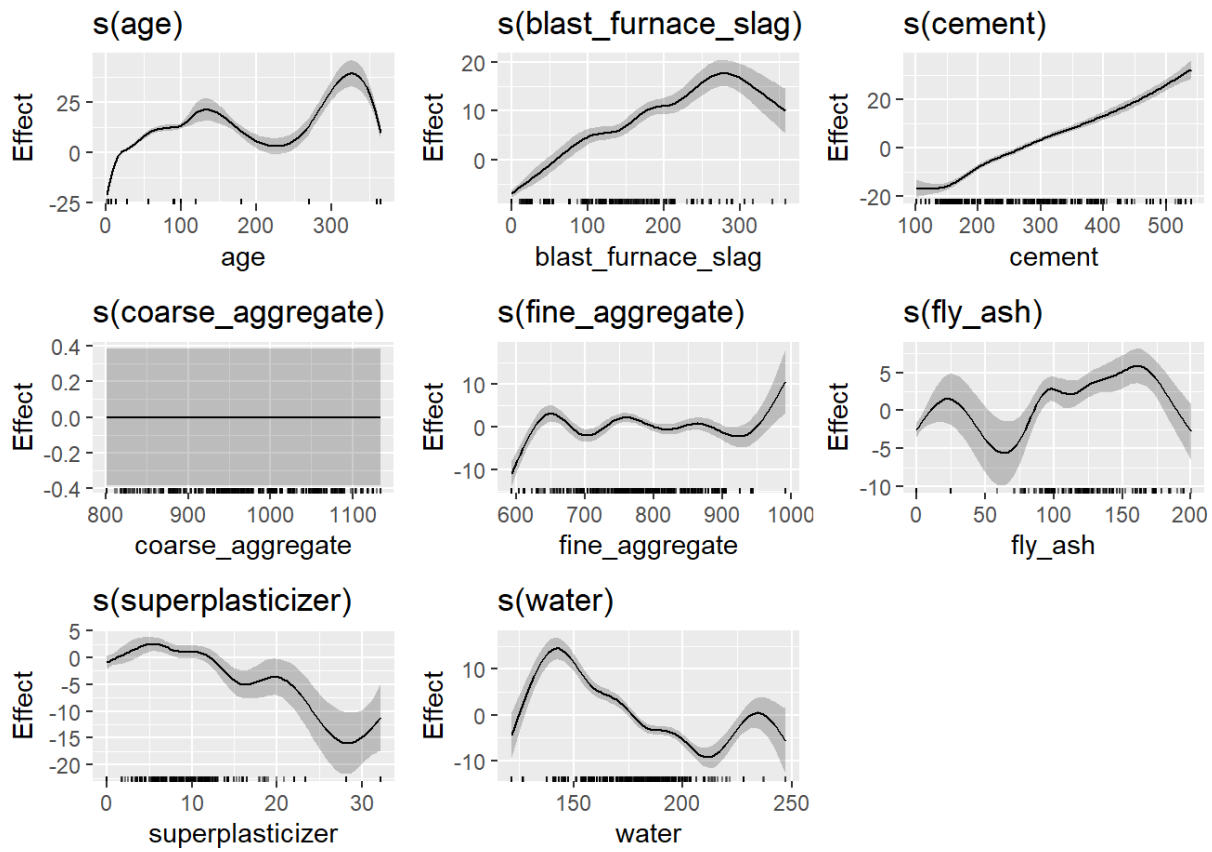

```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 128 iterations.
```

```
## The RMS GCV score gradient at convergence was 0.0003962093 .
## The Hessian was not positive definite.
## Model rank =  73 / 73
##
## 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(cement)             9.00e+00 6.85e+00    0.91   0.020 *
## s(blast_furnace_slag) 9.00e+00 6.72e+00    0.96   0.085 .
## s(fly_ash)            9.00e+00 7.16e+00    1.01   0.655
## s(water)              9.00e+00 8.38e+00    0.94   0.005 **
## s(superplasticizer)   9.00e+00 7.85e+00    0.93   0.030 *
## s(coarse_aggregate)   9.00e+00 2.33e-07    0.86  <2e-16 ***
## s(fine_aggregate)     9.00e+00 8.51e+00    0.87  <2e-16 ***
## s(age)                9.00e+00 7.99e+00    1.06   0.950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(model_gam)
```

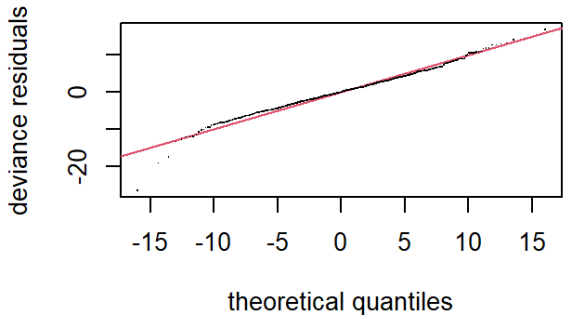
```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## strength ~ s(cement) + s(blast_furnace_slag) + s(fly_ash) + s(water) +
##     s(superplasticizer) + s(coarse_aggregate) + s(fine_aggregate) +
##     s(age)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.7507     0.1955   177.7   <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(cement)             6.853e+00      9 100.841  <2e-16 ***
## s(blast_furnace_slag) 6.723e+00      9  50.595  <2e-16 ***
## s(fly_ash)            7.162e+00      9   6.798  <2e-16 ***
## s(water)              8.383e+00      9  29.612  <2e-16 ***
## s(superplasticizer)   7.851e+00      9   7.709  <2e-16 ***
## s(coarse_aggregate)   2.332e-07      9   0.000   0.958
## s(fine_aggregate)     8.514e+00      9   9.957  <2e-16 ***
## s(age)                7.986e+00      9 266.120  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.887   Deviance explained = 89.5%
## GCV = 31.031  Scale est. = 28.787    n = 753
```

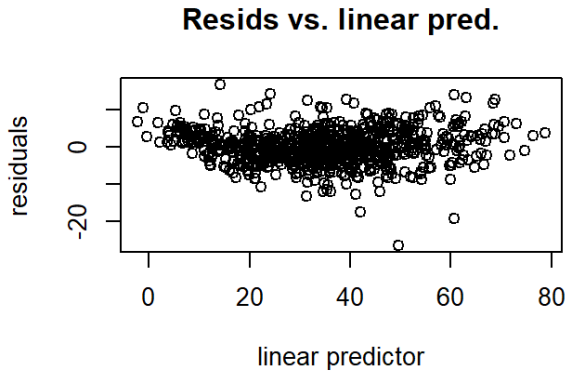
```
draw(model_gam)
```

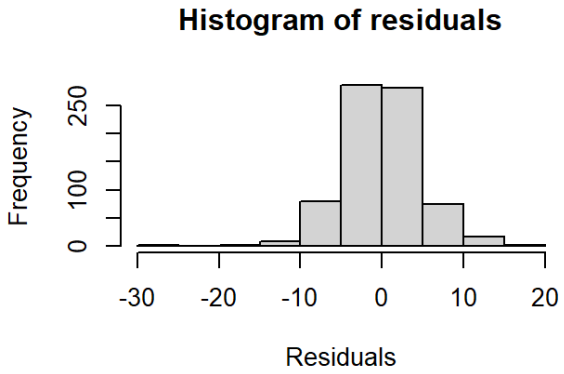


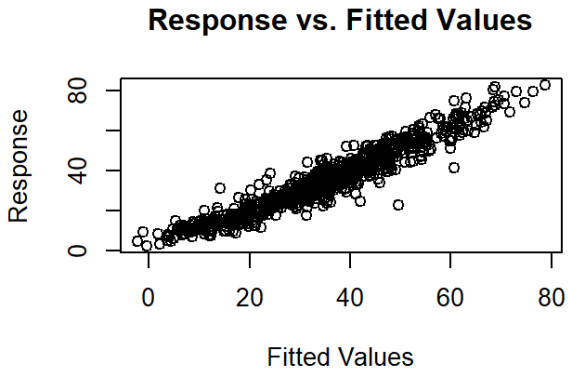
Padidinus mazgų skaičių ir naudojant visas prieš tai aprašytas diagnostikas nerasta stiprių nukrypimų nuo teorinio modelio. Gautame modelyje visos kovariantės reikšmingos, taip pat gautas statistiškai reikšmingas skirtumas nuo modelio, naudojančio mažesnį (numatytąjį) mazgų skaičių ($p<0.001$). Siekiant palyginti prieš tai sudaryta daug paprastenį polinominės regresijos ir gautą glodniųjų splainų modelį, abiems iš jų nubraižyti kovariančių efektų grafikai. Matoma, kad šiuo atveju gauti daug sudėtingesni sąryšiai tarp kovariančių ir atsako.

```
# Padidinamas mazgų skaičius
model_gam2 <- gam(strength ~ s(cement) + s(blast_furnace_slag, k = 20) + s(fly_ash, k = 20)
  + s(water, k = 20) + s(superplasticizer, k = 20)
  + s(coarse_aggregate, k = 20) + s(fine_aggregate, k = 20) + s(age, k = 10),
data = concrete_train,
select = TRUE
)

gam.check(model_gam2)
```



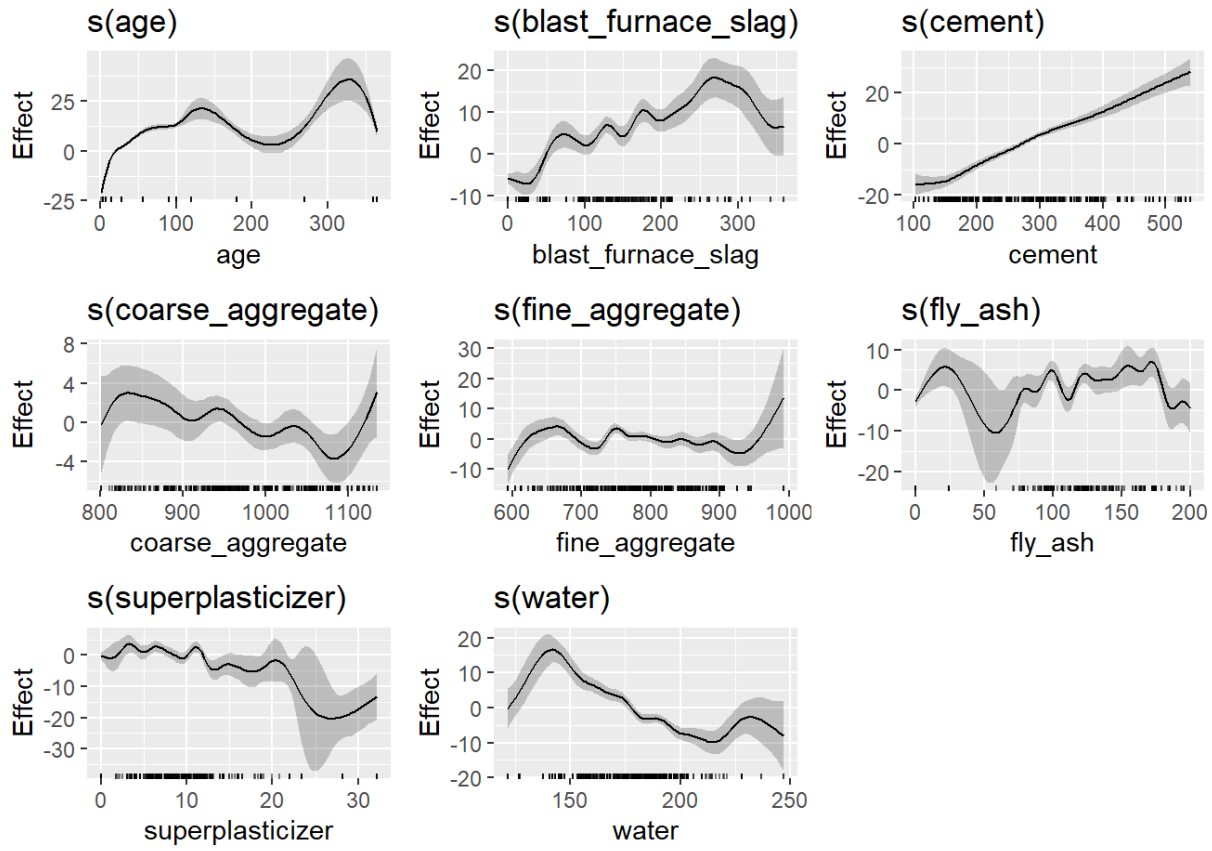




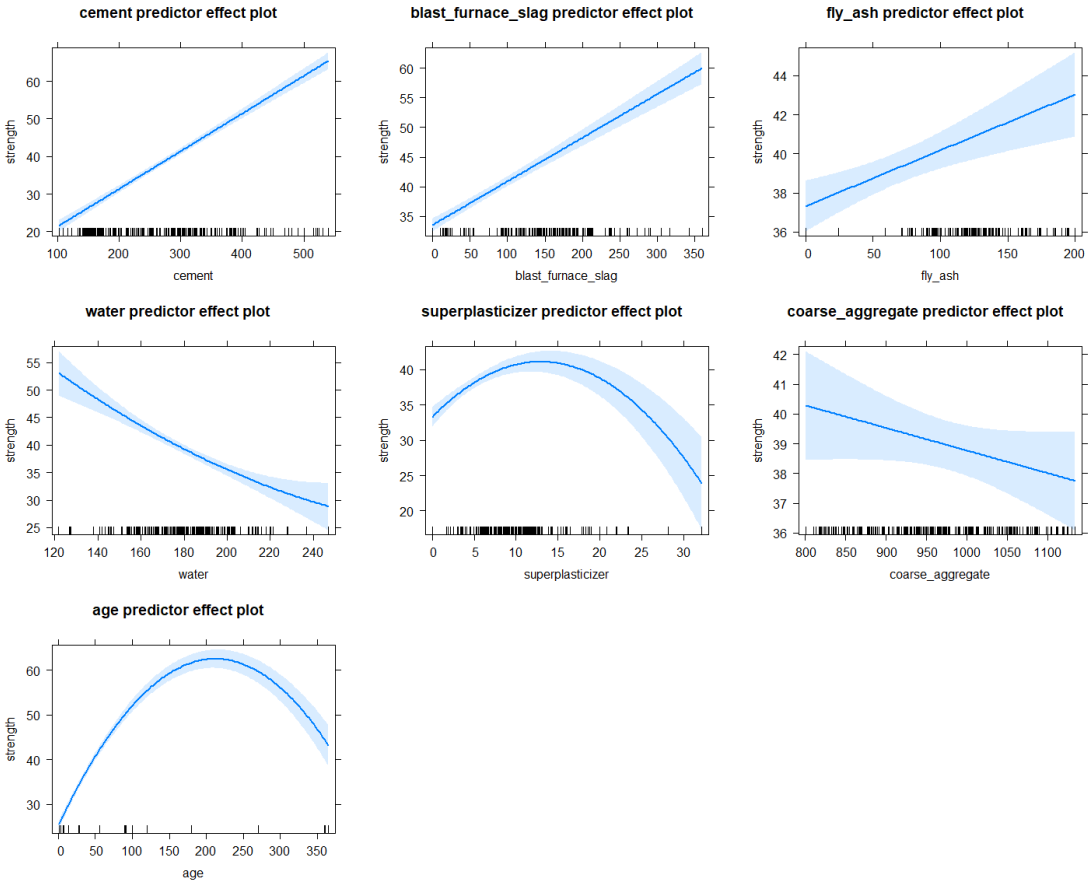



```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 82 iterations.
## The RMS GCV score gradient at convergence was 0.0004353583 .
## The Hessian was not positive definite.
## Model rank =  133 / 133
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##                          k'   edf k-index p-value
## s(cement)              9.00  6.19    0.99   0.375
## s(blast_furnace_slag) 19.00 13.96    1.03   0.845
## s(fly_ash)            19.00 15.09    1.05   0.940
## s(water)              19.00 12.16    1.01   0.635
## s(superplasticizer)   19.00 16.00    0.97   0.220
## s(coarse_aggregate)   19.00  9.28    0.96   0.105
## s(fine_aggregate)     19.00 15.01    0.95   0.055 .
## s(age)                 9.00  8.18    1.07   0.990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model_gam2)

##
## Family: gaussian
## Link function: identity
##
## Formula:
```

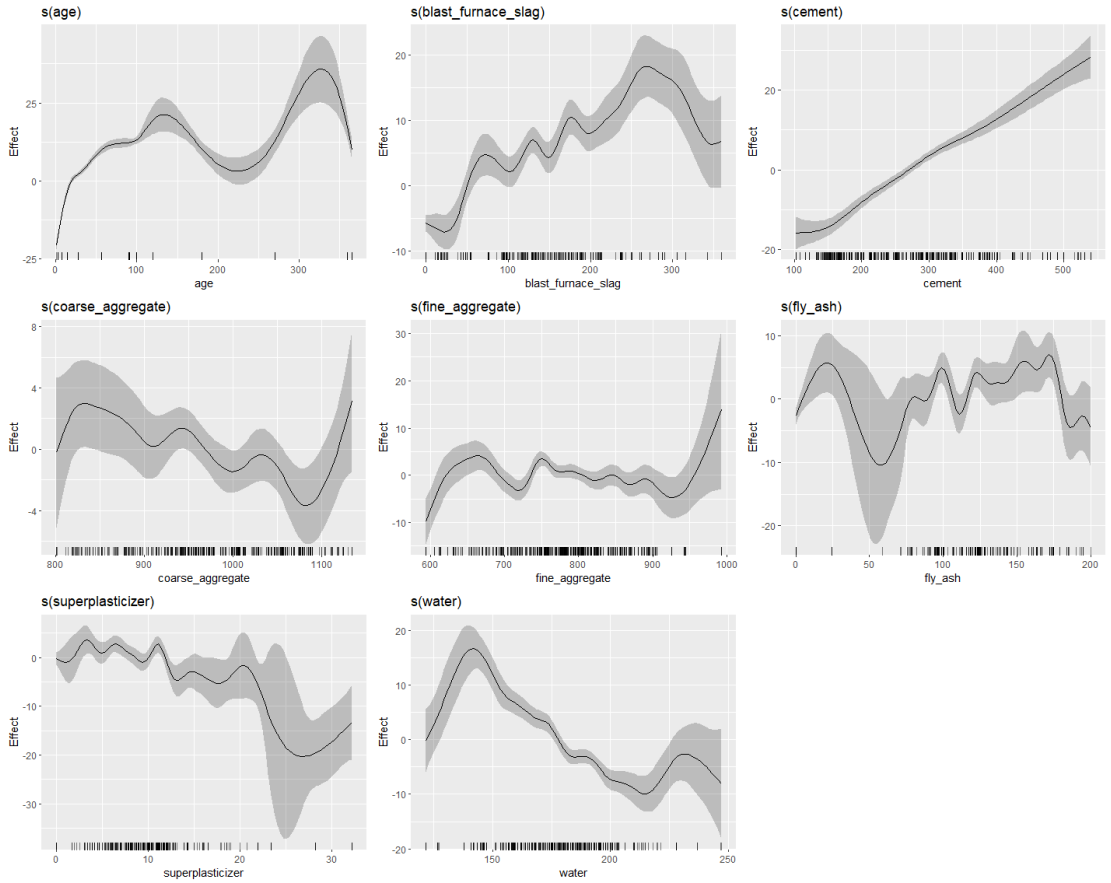
```
## strength ~ s(cement) + s(blast_furnace_slag, k = 20) + s(fly_ash,
##     k = 20) + s(water, k = 20) + s(superplasticizer, k = 20) +
##     s(coarse_aggregate, k = 20) + s(fine_aggregate, k = 20) +
##     s(age, k = 10)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  34.7507     0.1821   190.9   <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(cement)               6.188      9  23.149 < 2e-16 ***
## s(blast_furnace_slag) 13.959     19   7.039 < 2e-16 ***
## s(fly_ash)            15.091     19   3.670 < 2e-16 ***
## s(water)              12.163     19   8.275 < 2e-16 ***
## s(superplasticizer)   16.005     19   3.923 < 2e-16 ***
## s(coarse_aggregate)    9.276     19   1.151 0.00268 **
## s(fine_aggregate)     15.014     19   5.826 < 2e-16 ***
## s(age)                 8.177      9 298.256 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.902   Deviance explained = 91.5%
## GCV = 28.651  Scale est. = 24.965    n = 753

draw(model_gam2)
```



```r
# modelis statistiškai reikšmingai skiriasi nuo modelio su mažesniu mazgų skaičiumi
anova(model_gam, model_gam2,test="F")

## Analysis of Deviance Table
##
## Model 1: strength ~ s(cement) + s(blast_furnace_slag) + s(fly_ash) + s(water) +
##     s(superplasticizer) + s(coarse_aggregate) + s(fine_aggregate) +
##     s(age)
## Model 2: strength ~ s(cement) + s(blast_furnace_slag, k = 20) + s(fly_ash,
##     k = 20) + s(water, k = 20) + s(superplasticizer, k = 20) +
##     s(coarse_aggregate, k = 20) + s(fine_aggregate, k = 20) +
##     s(age, k = 10)
##   Resid. Df Resid. Dev     Df Deviance      F   Pr(>F)
## 1    694.30      20108
## 2    643.57      16380 50.732   3728.2 2.9437 3.48e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# modelio palyginimas su prieš tai sudarytu paprastos regresijos modeliu
library(effects)
AIC(model_polynomial)

## [1] 5247.325

AIC(model_gam2)

## [1] 4651.725

plot(predictorEffects(model_polynomial))
```



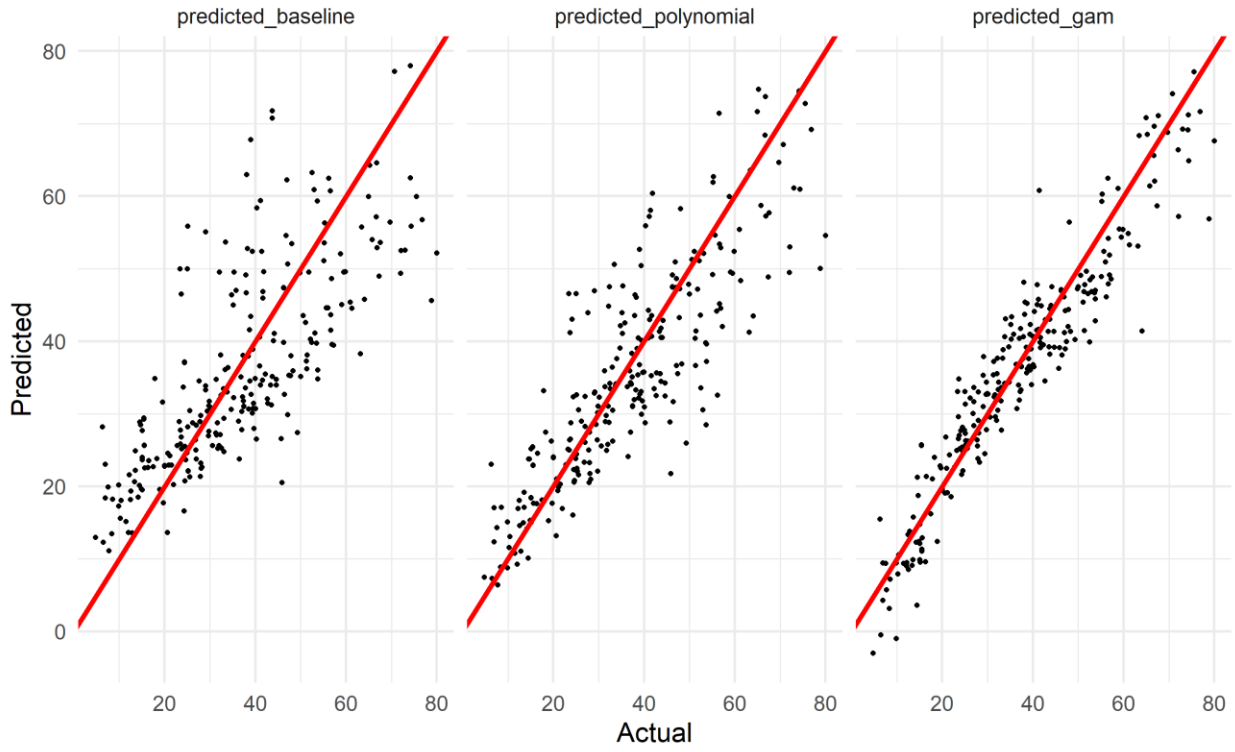

```
draw(model_gam2)
```



Kaip pastebėta naudojantis praėjusiais grafikais, glodniųjų splainų modeliu gauti daug sudėtingesni sąryšiai negu paprastu polinominės regresijos. Nors naudoti glodnieji splainai pasirinkta panaudoti testavimo aibę patikrinti, ar modelis tikrai nepersimokė ir įvertinti visų sudarytų modelių prognozavimo kokybę. Gauta, kad tiek pagal vidutinę absoliučią paklaidą (Mean Absolute Error), tiek matuojant vidutinę kvadratinę paklaidą ((Root) Mean Square Error), glodniųjų splainų modeliu gauti žymiai geresni rezultatai (paklaidos daugiau nei dvigubai mažesnės lyginant su paprastu tiesiniu modeliu, apie 50% mažesnės lyginant su polinominiu modeliu). Grafiškai modelių prognozės palygintos naudojant prognozuotų ir tikrų reikšmių sklaidos diagramas. Matoma, kad glodniųjų splainų modelių gauta daug mažesnė sklaida aplink prognozuotos ir tikros reikšmės lygybės tiesę lyginant su kitais modeliais.

```
# Nors naudoti glodnieji splainai, naudojant testavimo aibę patikrinama ar kažkur nebuvo
padaryta klaidų ir
# modelis tikrai nepersimokė
library(yardstick)

concrete_test <- concrete_test %>%
  mutate(predicted_baseline = predict(baseline, concrete_test),
         predicted_polynomial = predict(model_polynomial,concrete_test),
         predicted_gam = predict(model_gam2,concrete_test))

set <- metric_set(rmse,mae)
set(concrete_test, strength, predicted_baseline)
```

```
## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard       11.1
## 2 mae     standard        8.79

set(concrete_test, strength, predicted_polynomial)

## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        9.08
## 2 mae     standard        6.76

set(concrete_test, strength, predicted_gam)

## # A tibble: 2 x 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 rmse    standard        5.47
## 2 mae     standard        4.28

# tiek pagal MAE, tiek pagal RMSE naudojant gam modelį gaunami geresni rezultatai


# Modelių prognozių palyginimas
concrete_test %>%
  pivot_longer(c(predicted_gam, predicted_polynomial,predicted_baseline)) %>%
  ggplot(aes(strength, value)) +
  geom_point() +
  facet_wrap(vars(name)) +
  geom_abline(color = "red", size = 1.25) +
  labs(x = "Actual", y = "Predicted") +
  theme_minimal()
```



Išvados:

Atlikus pirminę duomenų aibės analizę rasta, kad betono stipris gali netiesiškai priklausyti nuo jo mišinį sudarančių medžiagų, todėl laikyta, kad paprastas tiesinės regresijos modelis gali būti netinkamas prognozuoti betono stiprį.

Siekiant turėti palyginamąjį modelį sudarytas paprastas tiesinės regresijos modelis. Kaip vienas iš būdų pagerinti tiesinį modelį buvo pasirinkta polinominė regresija į modelį įtraukiant antruosius laipsnius tų kintamųjų, kuriuose pagal diagnostinius grafikus rasta tiesinės regresijos modelio nepanaudotos informacijos. Pasitelkiant modelių grafines modelių diagnostikas ir statistinius testus antru modeliu gauti geresni rezultatai.

Paskutinis sudarytas apibendrintas adityvus regresijos modelis, naudojantis glodniuosius splainus. Parinkus didesnį mazgų skaičių modelio diagnostikose beveik nerasta nukrypimų. Visos modelyje naudotos kovariantės modelyje statistiškai reikšmingos.

Testavimo aibė panaudota įvertinti modelių gebėjimą prognozuoti reikšmes. Geriausi rezultatai gauti naudojant glodniųjų splainų modelį (paklaidos daugiau nei dvigubai mažesnės už tiesinį ir apie 50% mažesnės už polinominį modelį), todėl daroma išvada, kad šis modelis yra labiau tinkamas prognozuoti betono stiprį, lyginant su kitais dviem.