Vilniaus Universitetas

Regresinė analizė

Laboratorinis darbas

Darbą atliko:

Vainius Gataveckas, Matas Gaulia, Dovydas Martinkus

Duomenų Mokslas

3 kursas 2 gr.

Vilnius, 2021

Naudoti metodai

Darbas atliktas naudojant R, SAS ir Python.

Naudoti R paketai:

tidyverse
janitor
car
lmtest
RcmdrMisc
lm.beta
psych
ppcor

Duomenys ir jų šaltiniai

Šalių gyventojų vidutinė gyvenimo trukmė pagal sveikatos rodiklius.

Duomenų šaltinis - Kaggle. Prieiga per internetą: https://www.kaggle.com/kumarajarshi/life-expectancy-who

Originalus šaltinis – WHO. Prieiga per internetą: https://www.who.int/data/gho/data/indicators

2000-2015 metų 193 šalių duomenys. Duomenis sudaro šie stulpeliai:

„Country“ – šalis.
„Year“ – metai.
„Developed“ - šalies išsivystymo lygio kategorija.
„Life Expactancy“ – vidutinė gyvenimo trukmė šalyje.
„Adult Mortality“ - suaugusių mirtingumas (mirtys tarp 15 ir 60 metų 1000 gyventojų)
„Number of Infant Deaths“ – naujagimių mirtys 1000 gyventojų
„Alcohol“ – suvartojimas vienam gyventojui (gryno alkoholio litrais)
„Percentage Expenditure“ – išlaidos sveikatos apsaugai kaip procentas BVP vienam žmogui.
„Hepatitis B“ – imunizacija nuo hepatito B tarp 1 metų vaikų (proc.).
„Measles“ – imunizacija nuo tymų tarp 1 metų vaikų (proc.).
„BMI“ – vidutinis KMI visai šalies populiacijai.
„Under five deaths“ – mirtys iki 5 metų 1000 gyventojų
„Polio“ – imunizacija nuo poliomelito tarp 1 metų vaikų (proc.)
„Total expenditure“ – vyriausybės išlaidų sveikatos apsaugai dalis (proc.).
„Diphteria“ – imunizacija tarp 1 metų vaikų (proc.).
„HIV/AIDS“ – mirtys 1000 gimimų (nuo 0 iki 4 metų).
„GDP“ – BVP vienam žmogui (JAV doleriais).
„Population“ – Gyventojų kiekis.
„Thinness Age 10-19“ – plonumas tarp vaikų nuo 10 iki 19 metų (proc.).
„Thinness Age 5-10“ – plonumas tarp vaikų nuo 5 iki 9 metų (proc.).
„Income Composition of Resourses“ – Žmogaus socialinės raidos indeksas (HDI) ekonominiai kriteriai (nuo 0 iki 1).
„Schooling“ – Mokymosi metų kiekis (metais).

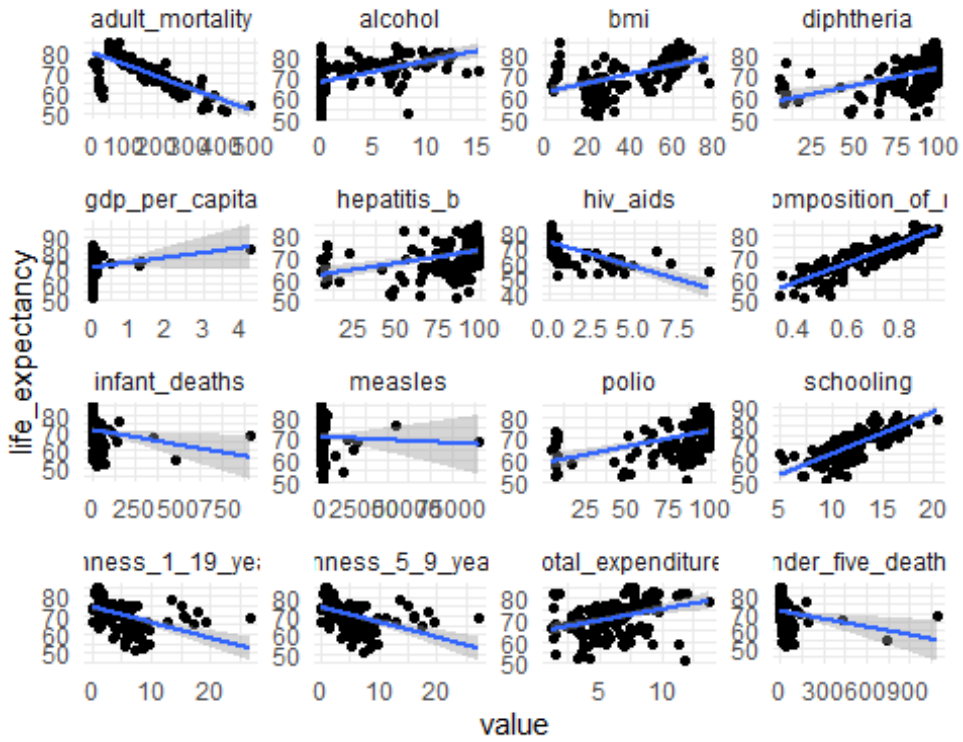
Atliktos analizės aprašymas

1. Naudojant R

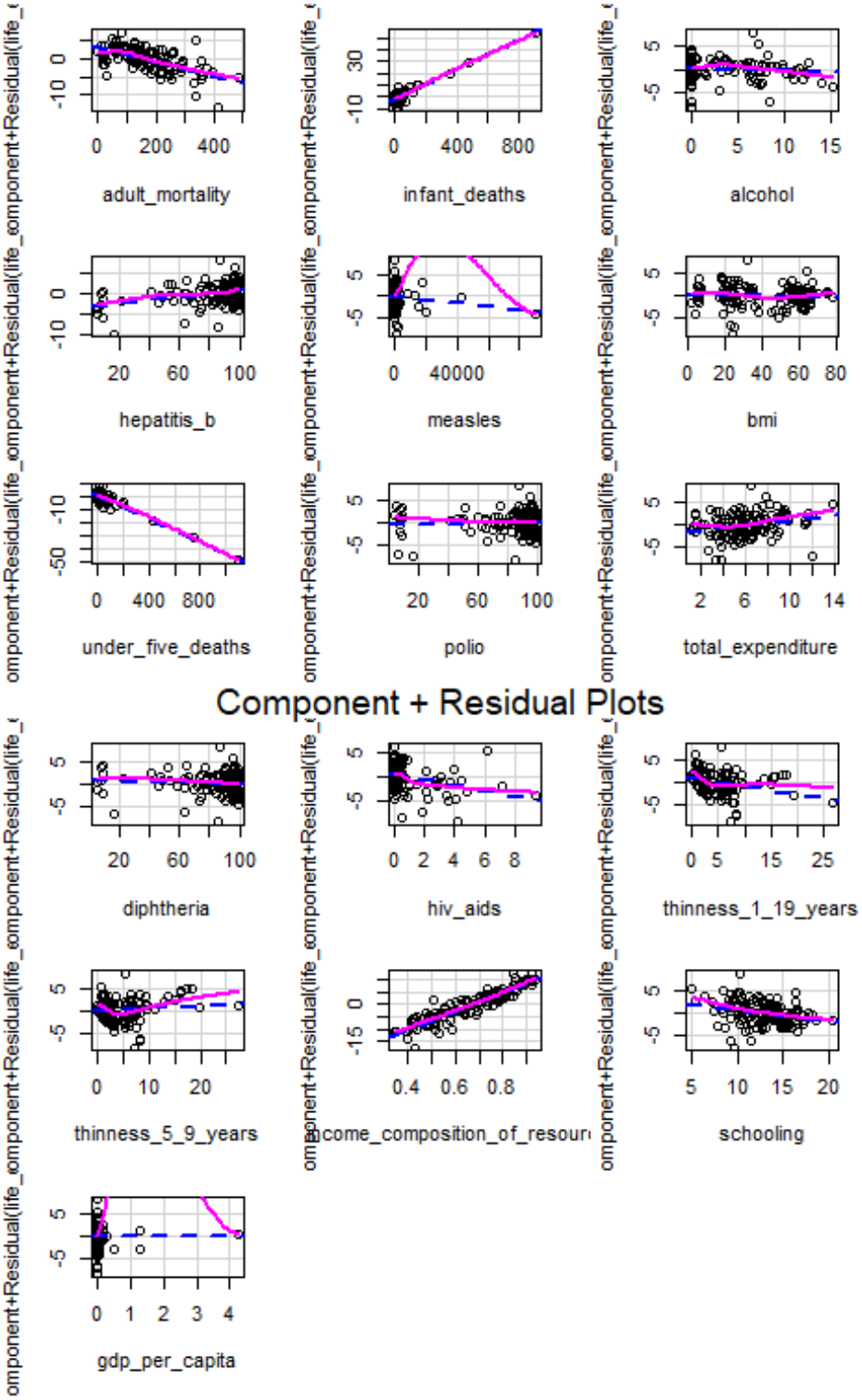
```
library(tidyverse)
library(car)
library(janitor)
x <- read_csv("life.csv") %>% clean_names()
```

Tikslas: prognozuoti vidutinę gyvenimo trukmę šalyje pagal tam tikrus sveikatos rodiklius.

```
set.seed(100)
transform_1<- function(x) {
  x %>%
  group_by(country) %>%
  fill(everything(), .direction = "up") %>%
  dplyr::select(-c(1, 3), -population, -percentage_expenditure) %>%
  drop_na() %>%
  ungroup() %>%
  dplyr::select(-1)
}

x <- transform_1(x)

x_1 <- x %>% filter(year == max(year)) %>% select(-1)

# atskiri duomenys, patikrinti kaip gautas galutinis modelis prognozuoja reikšmes
x_predict <- x %>% filter(year != max(year)) %>% slice_sample(n=10) %>% select(-1)


# kaikurių kovariančių priklausomybę nėra tiesinė
x_1 %>% pivot_longer(-1) %>% ggplot(aes(x=value,y=life_expectancy)) + facet_wrap(vars(name),sc
ales="free") + geom_point() + geom_smooth(method="lm") + theme_minimal()
```



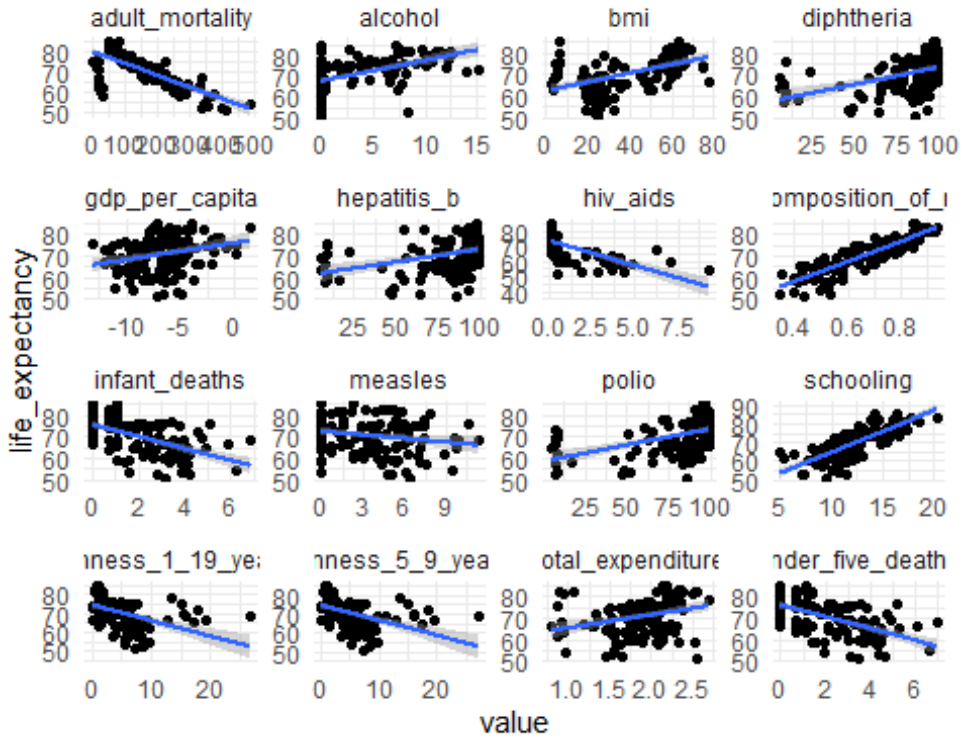

```
model <- lm(life_expectancy ~ ., data = x_1)
crPlots(model)
```



Rasta netiesinė priklausomybė tarp kai kurių kovariančių ir priklausomojo kintamojo. Kintamiesiems “gdp”, “infant_deaths”, “measles” ,”total_expenditure” ir “under_five_deaths” pastebėta stipri dešininė asimetrija (right skewedness), todėl pasirinkta atlikti log transformaciją.

```
transform_2 <- function(x) {
    x %>%
    mutate(gdp = log(gdp),
    infant_deaths = log(infant_deaths + 1),
    measles = log(measles + 1),
```

```
    total_expenditure = log(total_expenditure + 1),
    under_five_deaths = log(under_five_deaths + 1)
  )
}

# transformuojamos kaikurios kovariantės
x_2 <- transform_2(x_1)
x_predict <- transform_2(x_predict)


# Kintamųjų tiesinis ryšys patikrinamas dar kartą
x_2 %>% pivot_longer(-1) %>% ggplot(aes(x=value,y=life_expectancy)) + facet_wrap(vars(name),sc
ales="free") + geom_point() + geom_smooth(method="lm") + theme_minimal()
```

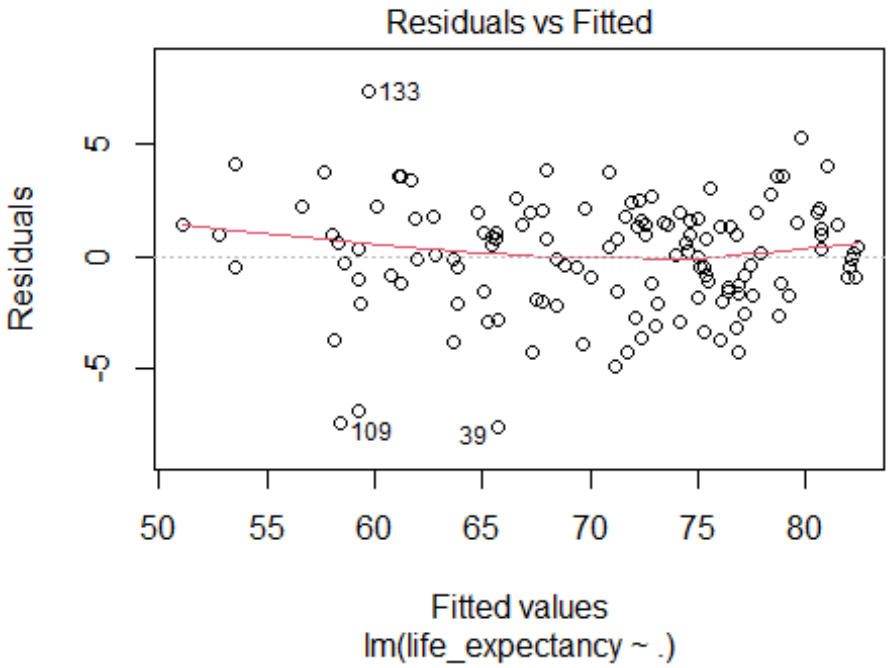


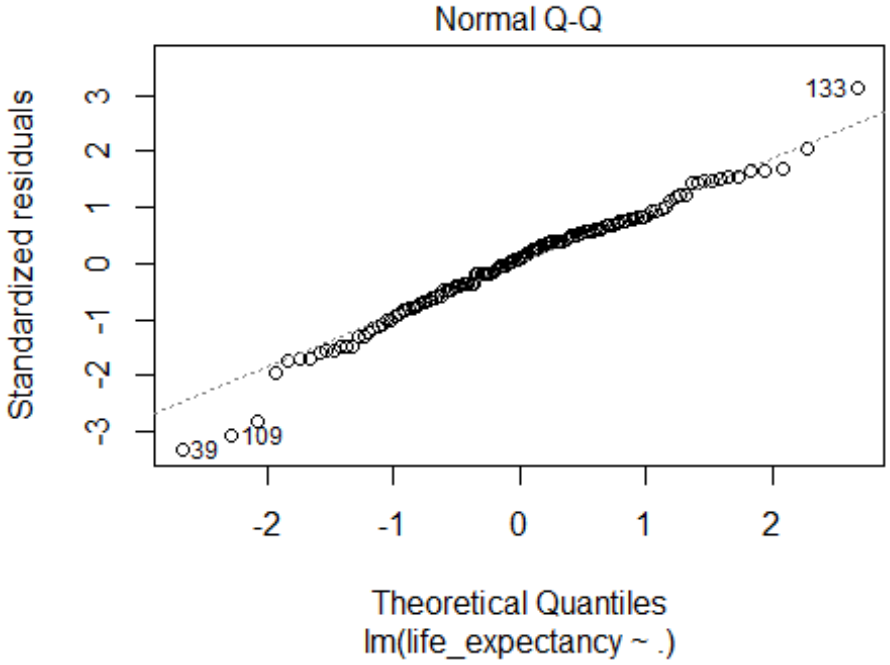
Modifikuoti duomenys išsaugomi faile „life_modified.csv“.

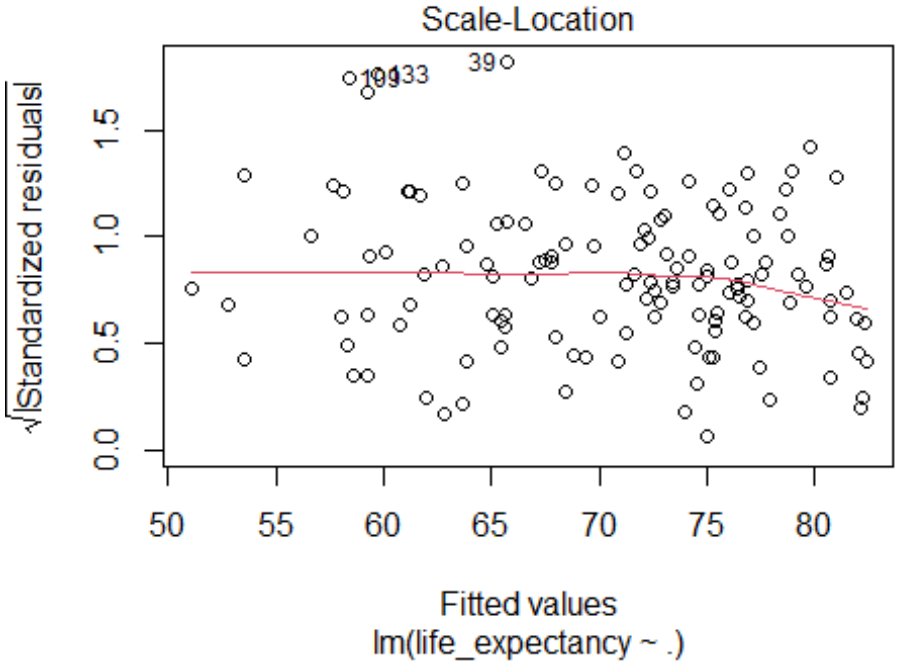
```
write.csv(x_2, "life_modified.csv")

# Sukuriamas modelis
model <- lm(life_expectancy ~ ., data = x_2)
```

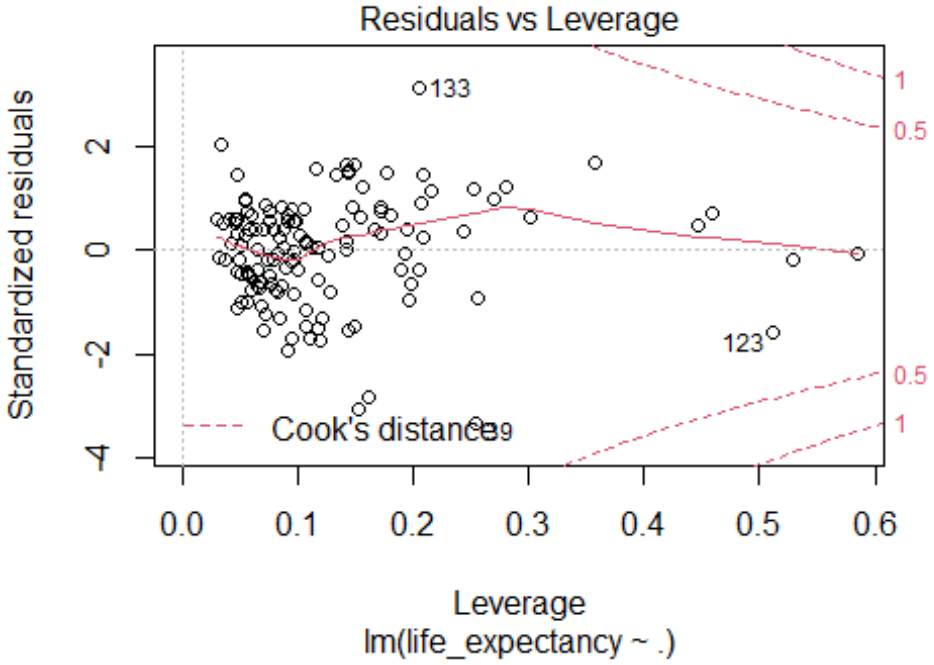
Modelio prielaidos

```
# Tikrinamas liekanų normalumas, homoskadiškumas, liekanų nepriklausomumas, išskirtys
plot(model)
```

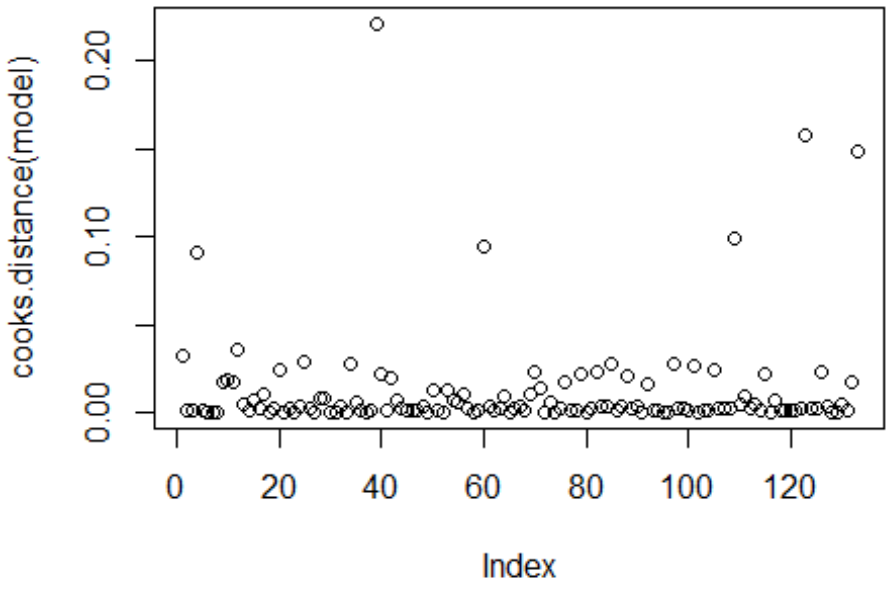

Residuals vs Fitted
Residuals
Fitted values
lm(life_expectancy ~ .)
133
109
39
50
55
60
65
70
75
80
5
0
-5


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(life_expectancy ~ .)
133
109
39
-2
-1
0
1
2
3

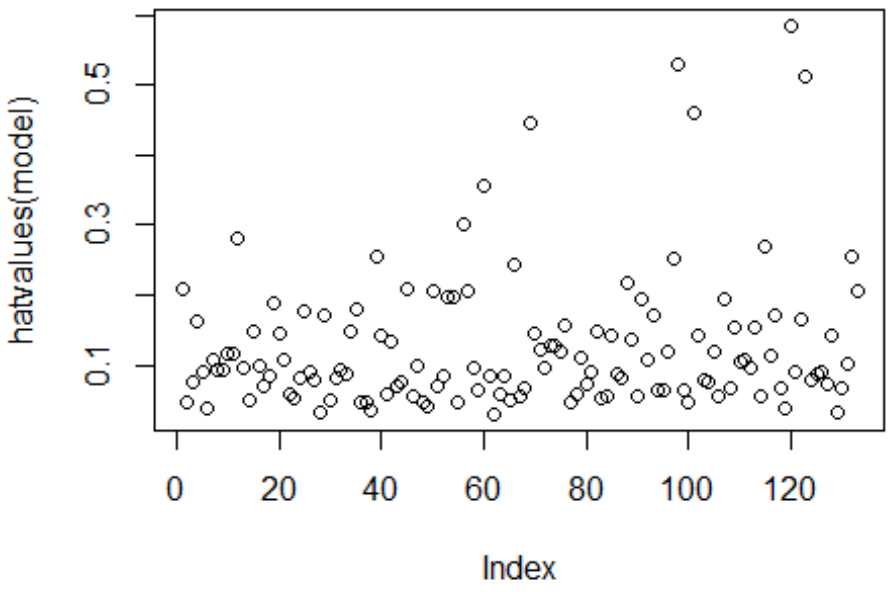





```
plot(cooks.distance(model))
```



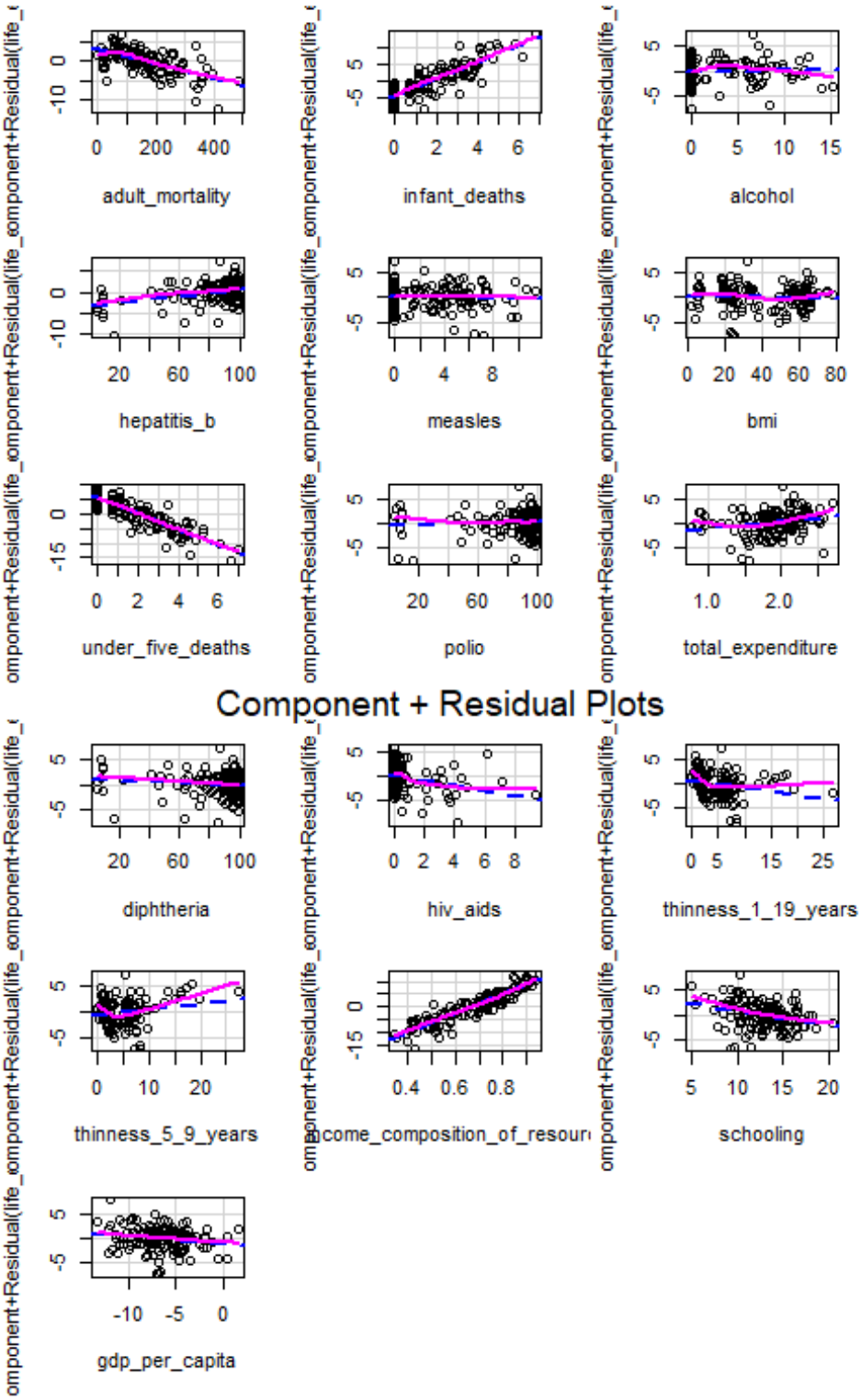

```
plot(hatvalues(model))
```




```
# Liekanų normalumo testas
shapiro.test(residuals(model))

##
##  Shapiro-Wilk normality test
##
## W = 0.98195, p-value = 0.07493
```

```r
# Homoskadiškumo testas
library(lmtest)
bptest(model)

## studentized Breusch-Pagan test
## BP = 13.511, df = 16, p-value = 0.6351

crPlots(model)
```

Tiek naudojant grafikus, tiek statistinius testus nerasta priklausomybės tarp liekanų, liekanų pasiskirstymo statistiško reikšmingo nuokrypio nuo normaliojo pasiskirstymo, išskirčių.




```
anova(model) # Tikrinama hipotezė H0: beta_1 = beta_2 = ... = 0

## Analysis of Variance Table
##
## Response: life_expectancy
##                                Df Sum Sq Mean Sq  F value    Pr(>F)
## adult_mortality                 1 4541.4  4541.4 658.0923 < 2.2e-16 ***
## infant_deaths                   1  714.3   714.3 103.5021 < 2.2e-16 ***
```

```
## alcohol                           1  631.9   631.9  91.5693 2.427e-16 ***
## hepatitis_b                       1  278.4   278.4  40.3488 4.305e-09 ***
## measles                           1    0.2     0.2   0.0300 0.8628941
## bmi                               1  152.7   152.7  22.1288 7.095e-06 ***
## under_five_deaths                 1  238.6   238.6  34.5813 4.022e-08 ***
## polio                             1   78.7    78.7  11.4067 0.0009967 ***
## total_expenditure                 1   33.3    33.3   4.8273 0.0300005 *
## diphtheria                        1    9.6     9.6   1.3904 0.2407448
## hiv_aids                          1   50.6    50.6   7.3376 0.0077755 **
## thinness_1_19_years               1   53.1    53.1   7.6883 0.0064776 **
## thinness_5_9_years                1    6.9     6.9   0.9952 0.3205464
## income_composition_of_resources   1  766.0   766.0 110.9948 < 2.2e-16 ***
## schooling                         1    9.0     9.0   1.3108 0.2546025
## gdp_per_capita                    1   19.2    19.2   2.7882 0.0976592 .
## Residuals                       116  800.5     6.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hipotezė apie reikšmingų kovariančių nebuvimą atmetama.

Modelio parinkimas

Parinkti modelį naudojama „backward/forward“ pažingsninė regresija. Išrenkamas modelis su 5 kovariantėmis.

```
# Požinksninė regresija
library(RcmdrMisc)
model_2 <- stepwise(model)

##
## Direction:  backward/forward
## Criterion:  BIC
##
## Step:  AIC=278.2
## life_expectancy ~ adult_mortality + hepatitis_b + total_expenditure +
##     hiv_aids + income_composition_of_resources
##
##                                  Df Sum of Sq     RSS    AIC
## <none>                                        863.91 278.20
## - total_expenditure               1     37.46  901.37 278.96
## + measles                         1     11.09  852.82 281.37
## + schooling                       1      8.38  855.52 281.79
## + thinness_1_19_years             1      8.26  855.65 281.81
## + under_five_deaths               1      6.98  856.93 282.01
## + gdp_per_capita                  1      6.83  857.08 282.04
## + thinness_5_9_years              1      5.20  858.71 282.29
## + infant_deaths                   1      5.00  858.90 282.32
## - hiv_aids                        1     61.54  925.45 282.46
## + polio                           1      2.30  861.60 282.74
## + alcohol                         1      2.23  861.68 282.75
## + bmi                             1      0.30  863.61 283.04
## + diphtheria                      1      0.17  863.73 283.06
## - hepatitis_b                     1     89.00  952.91 286.35
## - adult_mortality                 1    248.42 1112.32 306.92
## - income_composition_of_resources  1   2064.50 2928.40 435.67
```

Parametrų vertinimas ir interpretacija

```
# Koeficientai
summary(model_2)

##
## Call:
## lm(formula = life_expectancy ~ adult_mortality + hepatitis_b +
```

```
##     total_expenditure + hiv_aids + income_composition_of_resources,
##     data = x_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1512 -1.5507  0.2728  1.6248  8.3196
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    46.015816   1.879961  24.477  < 2e-16 ***
## adult_mortality                -0.019823   0.003280  -6.043 1.56e-08 ***
## hepatitis_b                     0.035768   0.009888   3.617 0.000428 ***
## total_expenditure               1.383667   0.589638   2.347 0.020491 *
## hiv_aids                       -0.608046   0.202160  -3.008 0.003174 **
## income_composition_of_resources 33.937181   1.948050  17.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.608 on 127 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8929
## F-statistic: 221.1 on 5 and 127 DF,  p-value: < 2.2e-16
```

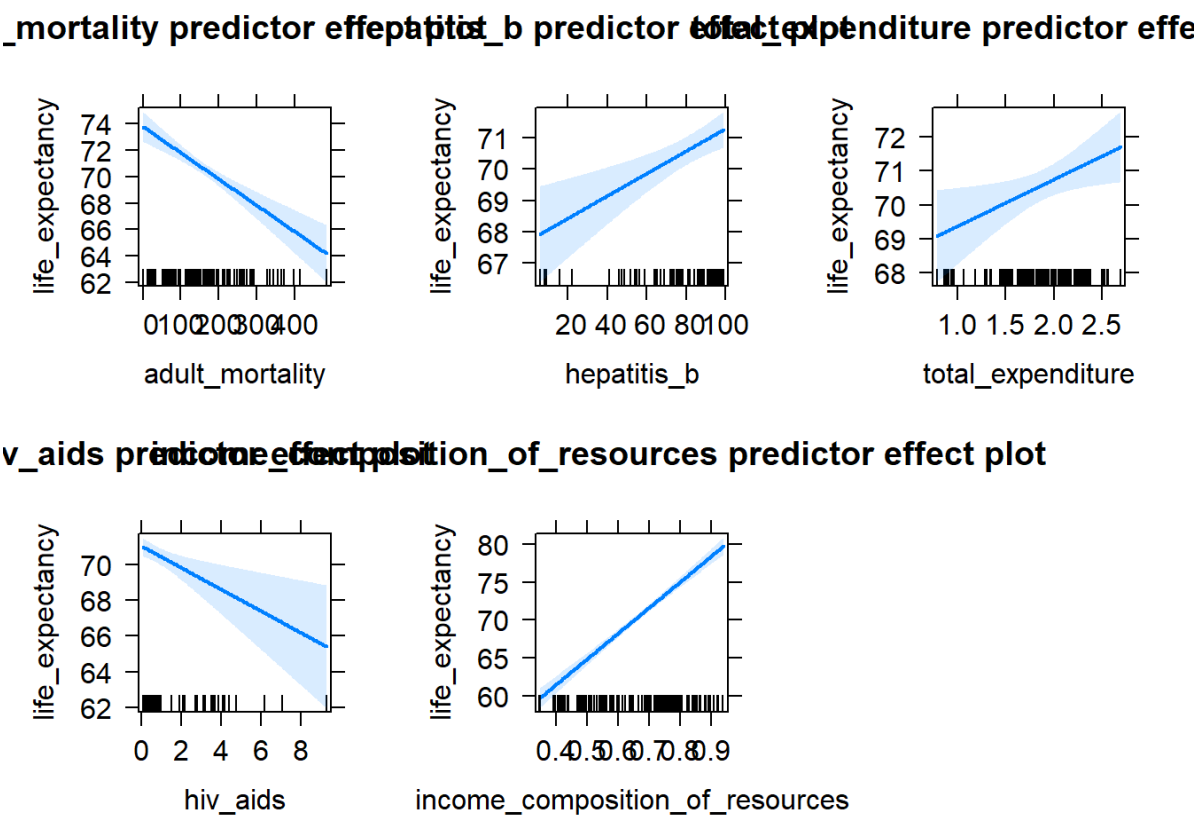
```
  # Visų koeficientų interpretacija paprasta,
  # nes pažingsnine regresija neišrinkti transformuoti kintamieji
library(lm.beta)
# Standartizuoti koeficientai
lm.beta(model_2)
```

```
##
## Call:
## lm(formula = life_expectancy ~ adult_mortality + hepatitis_b +
##     total_expenditure + hiv_aids + income_composition_of_resources,
##     data = x_2)
##
## Standardized Coefficients::
##                     (Intercept)                adult_mortality
##                      0.00000000                    -0.24840840
##                     hepatitis_b               total_expenditure
##                      0.11222105                     0.06927302
##                       hiv_aids income_composition_of_resources
##                     -0.11477877                     0.64768318
```

```
# Pasikliovimo interalai
confint(model_2)
```

```
##                                      2.5 %      97.5 %
## (Intercept)                    42.29571386 49.73591902
## adult_mortality                -0.02631364 -0.01333173
## hepatitis_b                     0.01620110  0.05533575
## total_expenditure               0.21687917  2.55045417
## hiv_aids                       -1.00808384 -0.20800885
## income_composition_of_resources 30.08234193 37.79202074
```

```
# Kovariančių įtaka vizualizuota
library(effects)
plot(predictorEffects(model_2))
```



Pažingsnine regresija parinktame modelyje tarp kovariančių nėra transformuotų kintamųjų, todėl visų koeficientų interpretacija įprasta.

Suaugusių mirtingumo (tikimybė mirti tarp 15 ir 60 metų 1000 gyventojų) (stulp. adult_*mortality*) ir mirčių nuo ŽIV/AIDS nuo 0 iki 4 metų 1000 gimimų (stulp. hiv_aids) didėjimas neigiamai įtakoja vidutinę gyvenimo trukmę.

Imunizacijos nuo Hepatito B tarp 1 metų vaikų % (stulp. hepatitis_b),
Dalies visų vyriausybės išlaidų sveikatos apsaugai (stulp. total_expenditure) ir
HDI pagal pajamų parametrą (stulp. income_composition_of_resources) didėjimas teigiamai įtakoja vidutinę gyvenimo trukmę.

Naudojant standartizuotus krypties koeficientus, didžiausia įtaką turinti kovariantė yra HDI pagal pajamų parametrą (stulp. income_composition_of_resources β=*0.65*), mažiausią - dalis visų vyriausybės išlaidų sveikatos apsaugai (stulp. total_expenditure β=*0.07*).

Multikolinearumo tikrinimas

```
vars <- dplyr::select(x_2, c(adult_mortality, hepatitis_b, total_expenditure,
  hiv_aids, income_composition_of_resources, life_expectancy))

#library(psych)
#corr.test(vars)

#dalinės koreliacijos
library(ppcor)
pcor(vars)$estimate
```

```
##                                adult_mortality  hepatitis_b total_expenditure
## adult_mortality                     1.00000000  0.284752689       0.031114658
## hepatitis_b                         0.28475269  1.000000000      -0.007076189
## total_expenditure                   0.03111466 -0.007076189       1.000000000
## hiv_aids                            0.30378653 -0.187990543       0.103610440
## income_composition_of_resources     0.18178399 -0.156298047      -0.086817301
## life_expectancy                    -0.47258053  0.305618694       0.203857631
##                                  hiv_aids income_composition_of_resources
## adult_mortality                 0.3037865                       0.1817840
## hepatitis_b                    -0.1879905                      -0.1562980
## total_expenditure               0.1036104                      -0.0868173
## hiv_aids                        1.0000000                       0.1721392
## income_composition_of_resources 0.1721392                       1.0000000
## life_expectancy                -0.2578685                       0.8396372
##                                life_expectancy
## adult_mortality                     -0.4725805
## hepatitis_b                          0.3056187
## total_expenditure                    0.2038576
## hiv_aids                            -0.2578685
## income_composition_of_resources      0.8396372
## life_expectancy                      1.0000000

# Variance inflation factor
vif(model_2)

##                 adult_mortality                     hepatitis_b
##                        2.082698                        1.186351
##               total_expenditure                        hiv_aids
##                        1.074114                        1.794951
## income_composition_of_resources
##                        1.703679
```

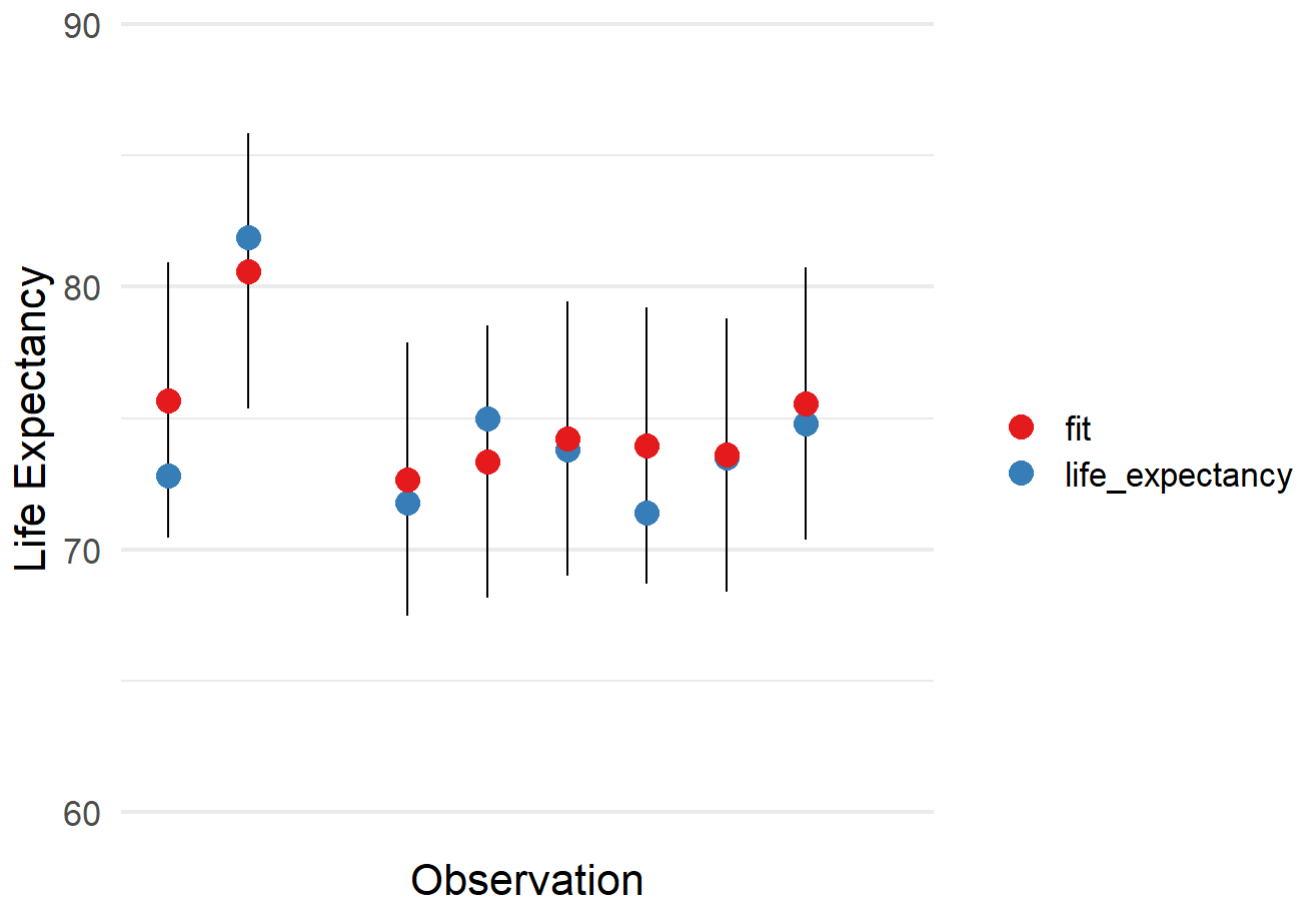
Naudojant dalinių koreliacijų matricą nerasta stiprių kovariančių tarpusavio koreliacijų. Variance inflation factor reiškmės <2.09 visoms modelyje esančioms kovariantėms.

Modelio tinkamumo analizė

```
summary(model_2)

##
## Call:
## lm(formula = life_expectancy ~ adult_mortality + hepatitis_b +
##     total_expenditure + hiv_aids + income_composition_of_resources,
##     data = x_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1512 -1.5507  0.2728  1.6248  8.3196
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     46.015816   1.879961  24.477  < 2e-16 ***
## adult_mortality                 -0.019823   0.003280  -6.043 1.56e-08 ***
## hepatitis_b                      0.035768   0.009888   3.617 0.000428 ***
## total_expenditure                1.383667   0.589638   2.347 0.020491 *
## hiv_aids                        -0.608046   0.202160  -3.008 0.003174 **
## income_composition_of_resources 33.937181   1.948050  17.421  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.608 on 127 degrees of freedom
## Multiple R-squared:  0.897,  Adjusted R-squared:  0.8929
## F-statistic: 221.1 on 5 and 127 DF,  p-value: < 2.2e-16
```

```r
  # R-squared = 0.897
  # Adj R-squared = 0.892

plot_predictions <- function(x,y) {
  predictions <- predict(x,newdata = y, interval = "prediction")
  predictions <- as_tibble(predictions) %>% mutate(n = 1:nrow(predictions))


  predictions_points <- y %>%
  mutate(pred = predictions) %>%
  unnest(pred) %>%
  dplyr::select(1,last_col(3),last_col(2),last_col(1),last_col(0)) %>%
  pivot_longer(c(1,2))


  ggplot(predictions) +
  geom_linerange(aes(x=n,ymin=lwr,ymax=upr)) +
  geom_point(data=predictions_points,aes(x=n,y=value,color=name),size = 4) +
  scale_x_discrete("Observation") +
  scale_y_continuous("Life Expectancy",limits = c(60,90)) +
  theme_minimal(base_size = 16) +
  scale_color_brewer("",palette = "Set1")
}

# Atliekamos kelios pavyzdinės prognozės
plot_predictions(model_2,x_predict)
```

Modelis paaiškina 89.7% duomenų sklaidos $R^2 = 0.897$. Modelio prognozės anksčiau nenaudotiems duomenims palyginamos su tikrosiomis vidutinės gyvenimo trukmės reikšmemis.



Rezultatai

Siekiant ištirti gyvenimo trukmės ryšį su sveikata susijusiais kriterijais naudota daugelio kintamųjų tiesinė regresija.

Pažingsnine regresija išrinktas modelis paaiškina 89.7% duomenų sklaidos (*$F(5,127) = 221.1$, $R^2 = 0.897$, $p<0.01$).* Rastos 5 statistiškai reikšmingos kovariantės gyvenimo trukmės prognozavimui (pateikti standartizuoti krypties koeficientai):

Suaugusių mirtingumas (tikimybė mirti tarp 15 ir 60 metų 1000 gyventojų) (stulp. adult_*mortality* *β=-0.25, p<0.001*)
Imunizacija nuo Hepatito B tarp 1 metų vaikų % (stulp. hepatitis_b *β=0.11, p<0.001*)
Dalis visų vyriausybės išlaidų sveikatos apsaugai (stulp. total_expenditure *β=0.07, p=0.02*)
Mirtys nuo ŽIV/AIDS nuo 0 iki 4 metų 1000 gimimų (stulp. hiv_aids *β=-0.11, p=0.003*)
HDI pagal pajamų parametrą (stulp. income_composition_of_resources *β=0.65, p<0.001*)

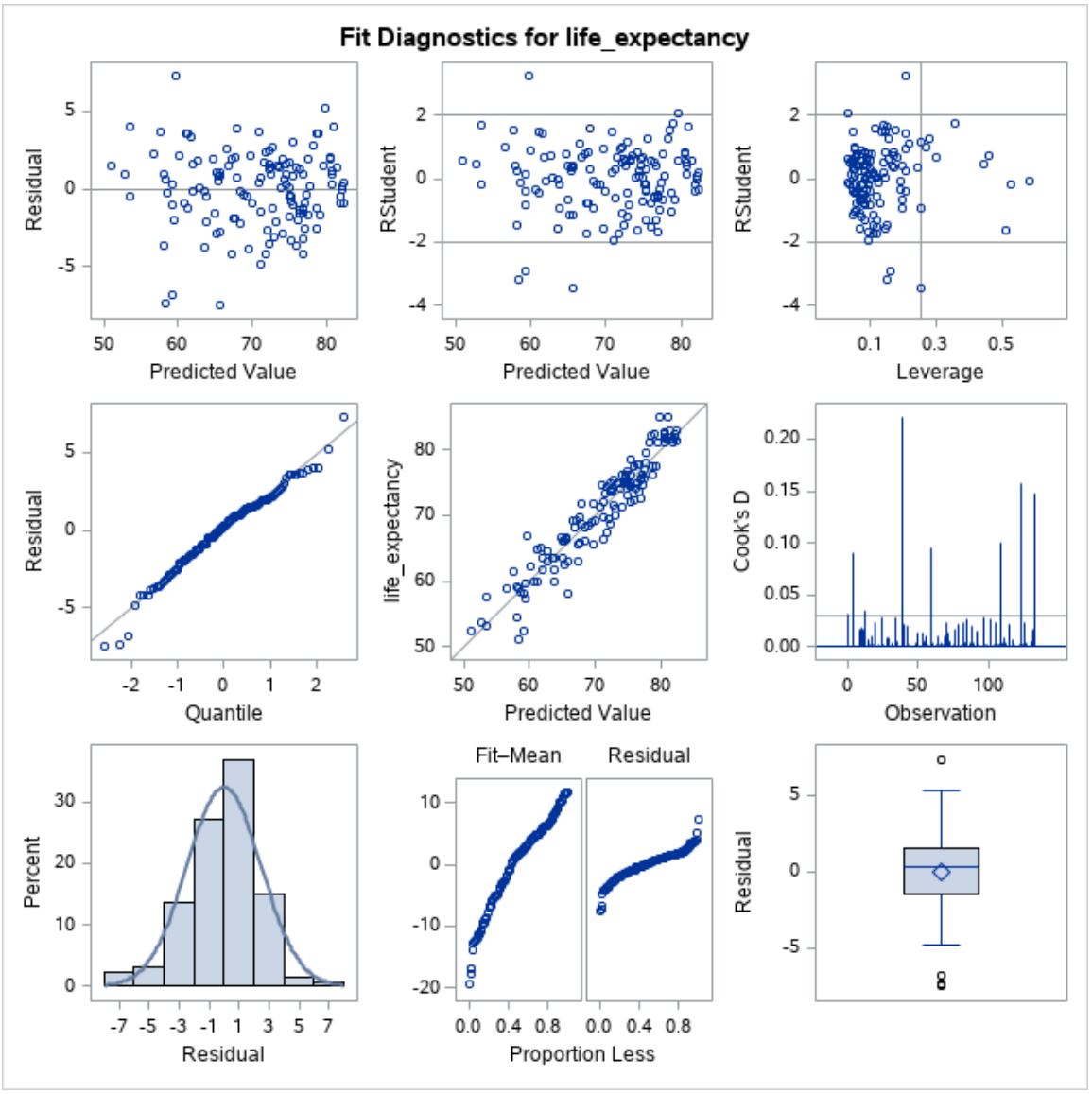
2. Naudojant SAS

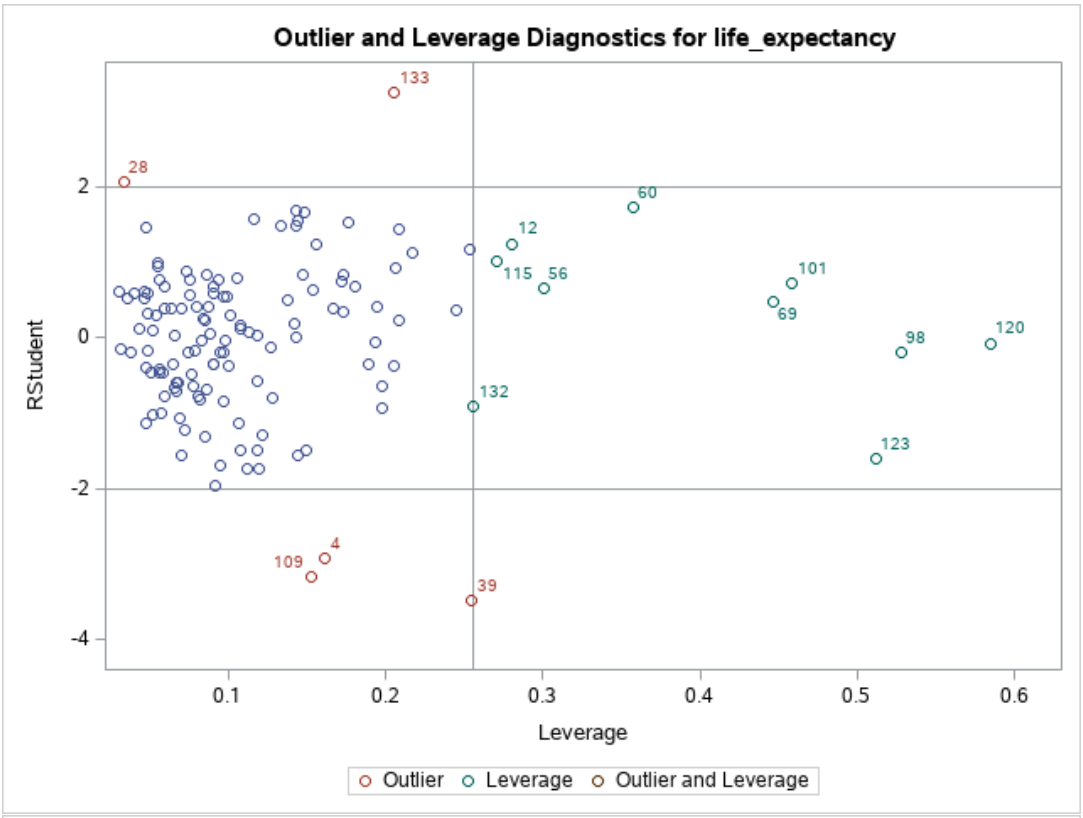
Naudojamas anksčiau sukurtas duomenų failas.

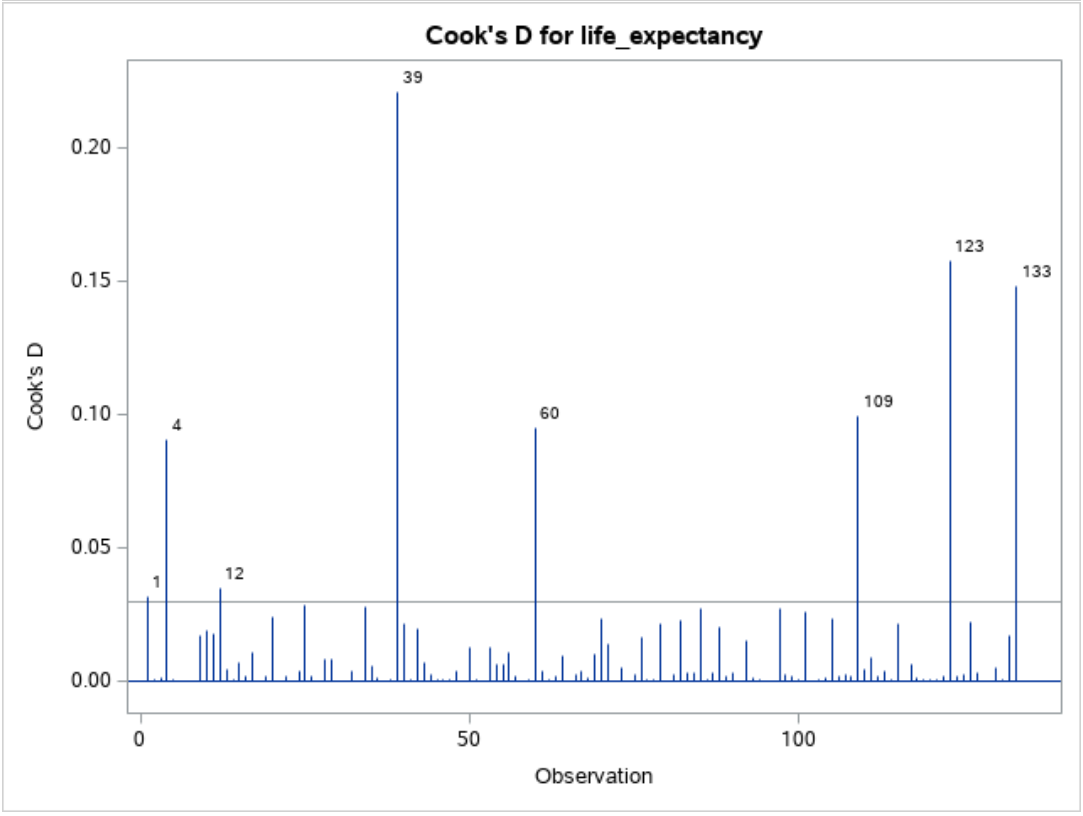
```
PROC IMPORT DATAFILE='/home/u45871880/life_modified.csv'
       DBMS=CSV
       OUT=data;
       GETNAMES=YES;
RUN;
```

Patikrinamos modelio prielaidos (liekanų normalumas, nepriklausomumas, homoskedastiškumas, išskirčių nebuvimas).

```
/* Modelio prielaidos */

PROC REG data=data simple corr plots=(diagnostics(stats=none) RStudentByLeverage(label)
             CooksD(label) Residuals(smooth) ObservedByPredicted(label));
MODEL life_expectancy = adult_mortality infant_deaths alcohol hepatitis_b measles
bmi under_five_deaths polio total_expenditure diphtheria hiv_aids
thinness_1_19_years thinness_5_9_years income_composition_of_resources
schooling gdp;
run;
```


Fit Diagnostics for life_expectancy
Residual
Predicted Value
RStudent
Predicted Value
RStudent
Leverage
Residual
Quantile
life_expectancy
Predicted Value
Cook's D
Observation
Percent
Residual
Fit–Mean
Residual
Proportion Less
Residual






```
/* Normalumo testas */
```

```
proc univariate data=rez normal;
var liekanos;
run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.981952	Pr < W	0.0749
Kolmogorov-Smirnov	D	0.060241	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.100101	Pr > W-Sq	0.1135
Anderson-Darling	A-Sq	0.63253	Pr > A-Sq	0.0979

```
/* Modelio parinkimas naudojant pažingsninę regresiją*/
/* Parametrų vertinimas */

PROC REG data=data plots=none outest=summary;
MODEL life_expectancy = adult_mortality infant_deaths alcohol hepatitis_b measles
bmi under_five_deaths polio total_expenditure diphtheria hiv_aids
thinness_1_19_years thinness_5_9_years income_composition_of_resources
schooling / stb vif cli clb pcorr2 slentry=0.05 slstay=0.05 selection=stepwise aic bic;
run;

proc print data=summary;
run;
```

Stepwise Selection: Step 6
Variable gdp Entered: R-Square = 0.9025 and C(p) = 5.2605

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	8131.77778	1355.29630	215.92	<.0001
Error	140	878.74195	6.27673		
Corrected Total	146	9010.51973			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	47.00391	1.85604	4025.57550	641.35	<.0001
adult_mortality	-0.01883	0.00311	230.10155	36.66	<.0001
hepatitis_b	0.03221	0.00940	73.78047	11.75	0.0008
total_expenditure	1.49427	0.53374	49.19615	7.84	0.0058
hiv_aids	-0.62505	0.19276	65.99640	10.51	0.0015
income_composition_of_resources	36.08633	2.23973	1629.39372	259.59	<.0001
gdp	-0.33851	0.16845	25.34869	4.04	0.0464

Bounds on condition number: 2.7001, 64.005

All variables left in the model are significant at the 0.0500 level.
No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	income_composition_of_resources		1	0.8138	0.8138	120.975	633.74	<.0001
2	adult_mortality		2	0.0568	0.8706	42.4851	63.17	<.0001
3	hepatitis_b		3	0.0157	0.8863	22.1786	19.79	<.0001
4	hiv_aids		4	0.0071	0.8934	14.0804	9.49	0.0025
5	total_expenditure		5	0.0062	0.8997	7.2488	8.75	0.0036
6	gdp		6	0.0028	0.9025	5.2605	4.04	0.0464

Matome, kad palyginus su užduoties atlikimu su R, pažingsninė regresija išrenka dar vieną papildomą kovariantę „gdp“.

3. Naudojant Python

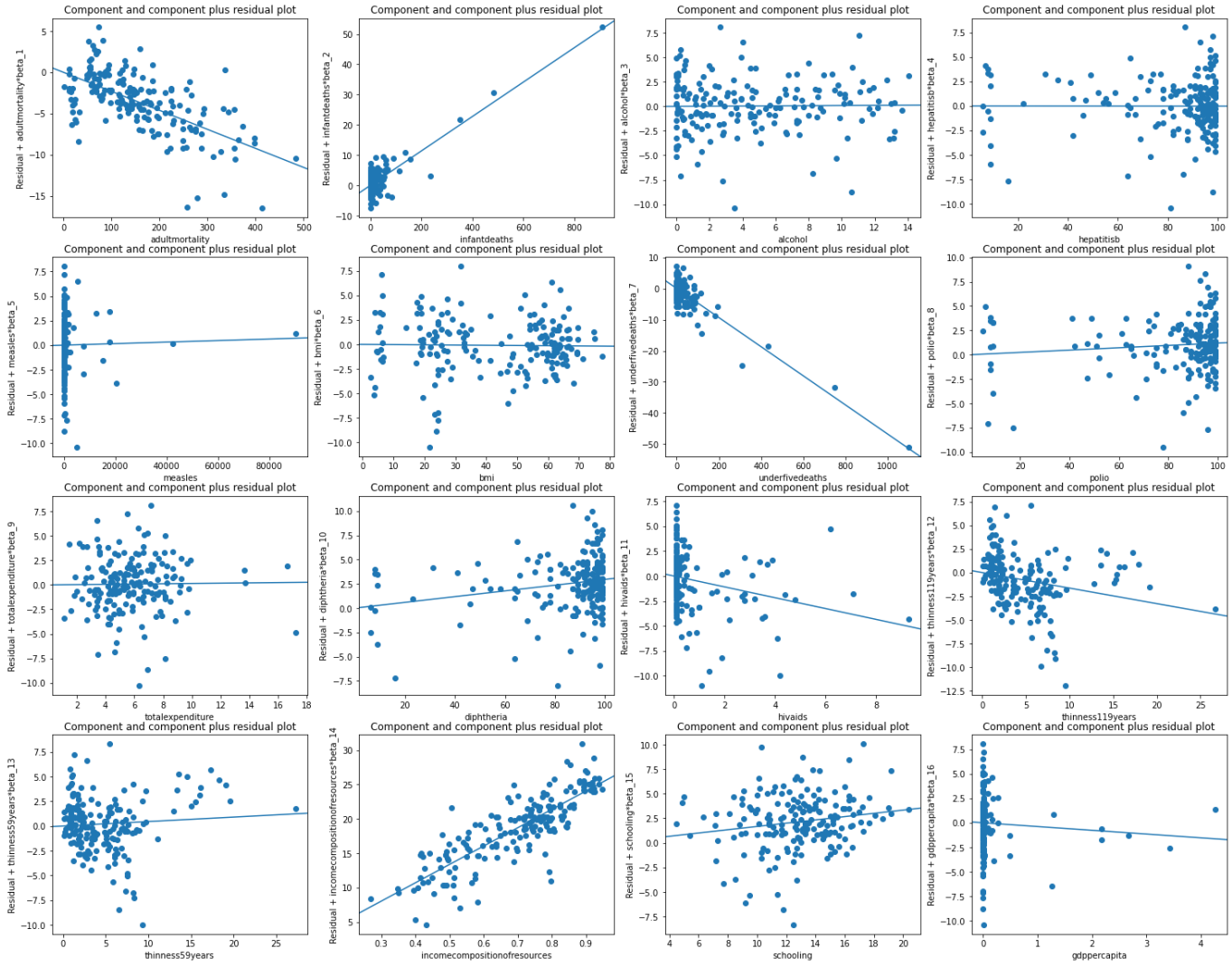
```
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
import statsmodels.api as sm
from statsmodels.formula.api import ols
from scipy import stats
from scipy.stats import shapiro
import statsmodels.stats.api as sms
from statsmodels.compat import lzip
def plot_for_every_column(model, columns):
    for c in columns:
        #fig = plt.figure(figsize=(12,8))
        #fig = sm.graphics.plot_regress_exog(model, c, fig=fig)
        fig = sm.graphics.plot_ccpr(model, c)
        fig.tight_layout(pad=1.0)

def plot_ccpr(model, cols):
    plotn = 0
    rows = 4
    columns = 4
    fig, ax_array = plt.subplots(rows, columns,squeeze=False)
    fig.set_figheight(20)
    fig.set_figwidth(25)
    for i,ax_row in enumerate(ax_array):
        for j,axes in enumerate(ax_row):
            axes.set_title(cols[plotn])
            sm.graphics.plot_ccpr(model, cols[plotn], ax = axes)
            plotn = plotn + 1
    plt.show()

def plot_model(df, model):
    influence = model.get_influence()

    df['resid'] = model.resid
    df['fittedvalues'] = model.fittedvalues
    df['resid_std'] = model.resid_pearson
    df['leverage'] = influence.hat_matrix_diag

    fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(15,8))
    plt.style.use('seaborn')

    # Residual against fitted values.
    df.plot.scatter(
        x='fittedvalues', y='resid', ax=axes[0, 0]
    )
    axes[0, 0].axhline(y=0, color='grey', linestyle='dashed')
    axes[0, 0].set_xlabel('Fitted Values')
    axes[0, 0].set_ylabel('Residuals')
    axes[0, 0].set_title('Residuals vs Fitted')

    # qqplot
    sm.qqplot(
        df['resid'], dist=stats.t, fit=True, line='45',
```

```
        ax=axes[0, 1], c='#4C72B0'
    )
    axes[0, 1].set_title('Normal Q-Q')

    # The scale-location plot.
    df.plot.scatter(
        x='fittedvalues', y='resid_std', ax=axes[1, 0]
    )
    axes[1, 0].axhline(y=0, color='grey', linestyle='dashed')
    axes[1, 0].set_xlabel('Fitted values')
    axes[1, 0].set_ylabel('Sqrt(|standardized residuals|)')
    axes[1, 0].set_title('Scale-Location')

    # Standardized residuals vs. leverage
    df.plot.scatter(
        x='leverage', y='resid_std', ax=axes[1, 1]
    )
    axes[1, 1].axhline(y=0, color='grey', linestyle='dashed')
    axes[1, 1].set_xlabel('Leverage')
    axes[1, 1].set_ylabel('Sqrt(|standardized residuals|)')
    axes[1, 1].set_title('Residuals vs Leverage')

    plt.tight_layout()
    plt.show()
d = pd.read_csv("life.csv")
d = d.interpolate(method = 'zero')
d["gdp_per_capita"] = d["GDP"] / d["Population"]
d.columns=d.columns.str.lower().str.replace(' ','')
d.columns=d.columns.str.lower().str.replace('-','')
d.columns=d.columns.str.lower().str.replace('/','')
d.columns=d.columns.str.lower().str.replace('_','')
d = d[d.year == max(d.year)]
d = d.drop(["country", "year","status", "gdp", "population",
"percentageexpenditure"], axis = 1)

f = "lifeexpectancy~" + "+".join(d.columns[1:])
Not normalised data
model = ols(formula = f, data=d).fit()
model.summary()
```

Dep. Variable:	lifeexpectancy	**R-squared:**	0.883
Model:	OLS	**Adj. R-squared:**	0.871
Method:	Least Squares	**F-statistic:**	78.10
Date:	Thu, 09 Dec 2021	**Prob (F-statistic):**	1.74e-68
Time:	19:55:53	**Log-Likelihood:**	-446.40
No. Observations:	183	**AIC:**	926.8
Df Residuals:	166	**BIC:**	981.4

Df Model:	16
Covariance Type:	nonrobust

```
plot_ccpr(model, d.columns[1:])
```



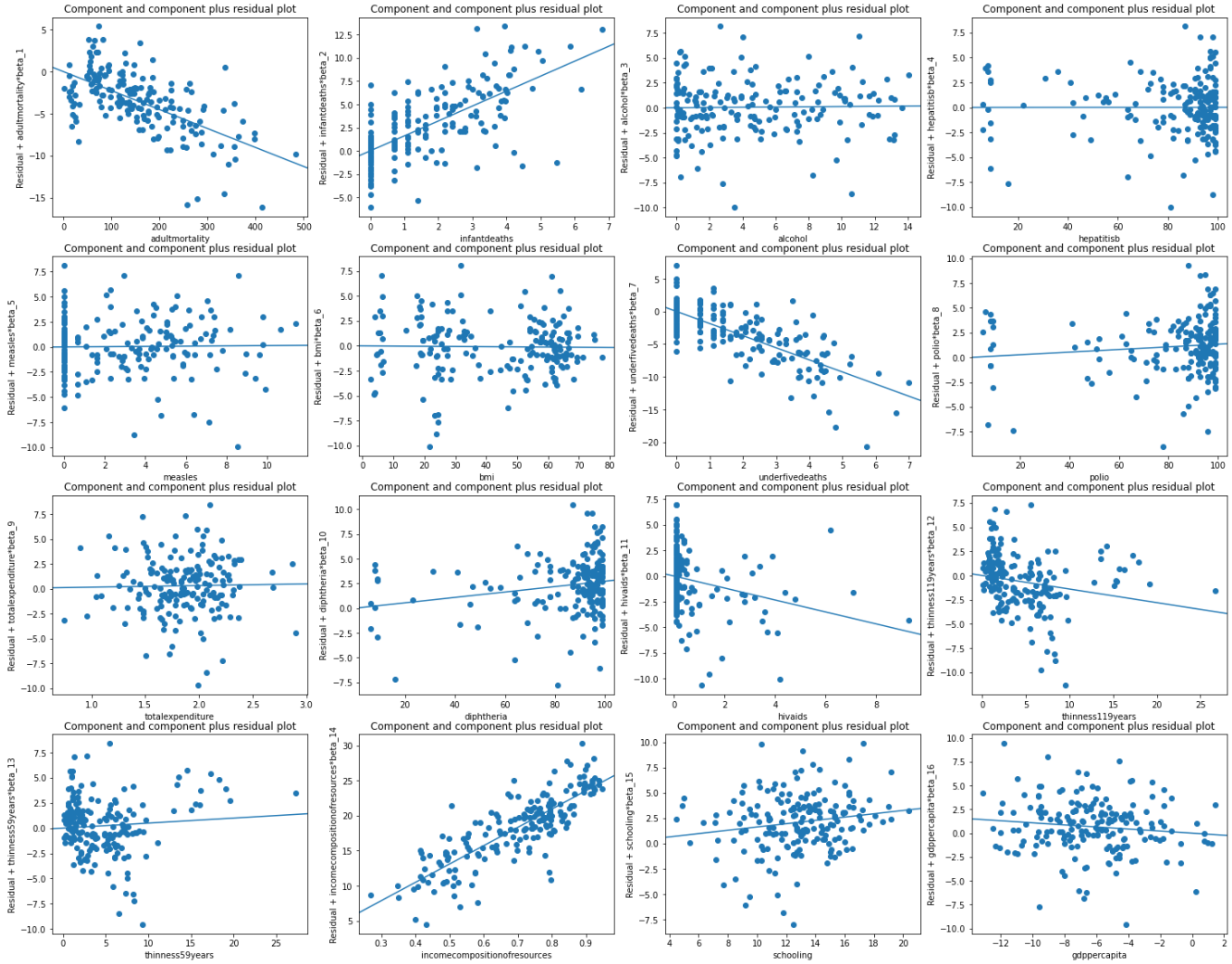
Normalised data

```
l = d.copy()
l.gdppercapita = np.log(l.gdppercapita)
l.infantdeaths = np.log(l.infantdeaths + 1)
l.measles = np.log(l.measles + 1)
l.totalexpenditure = np.log(l.totalexpenditure + 1)
l.underfivedeaths = np.log(l.underfivedeaths + 1)

model = ols(formula = f, data=l).fit()
model.summary()
```

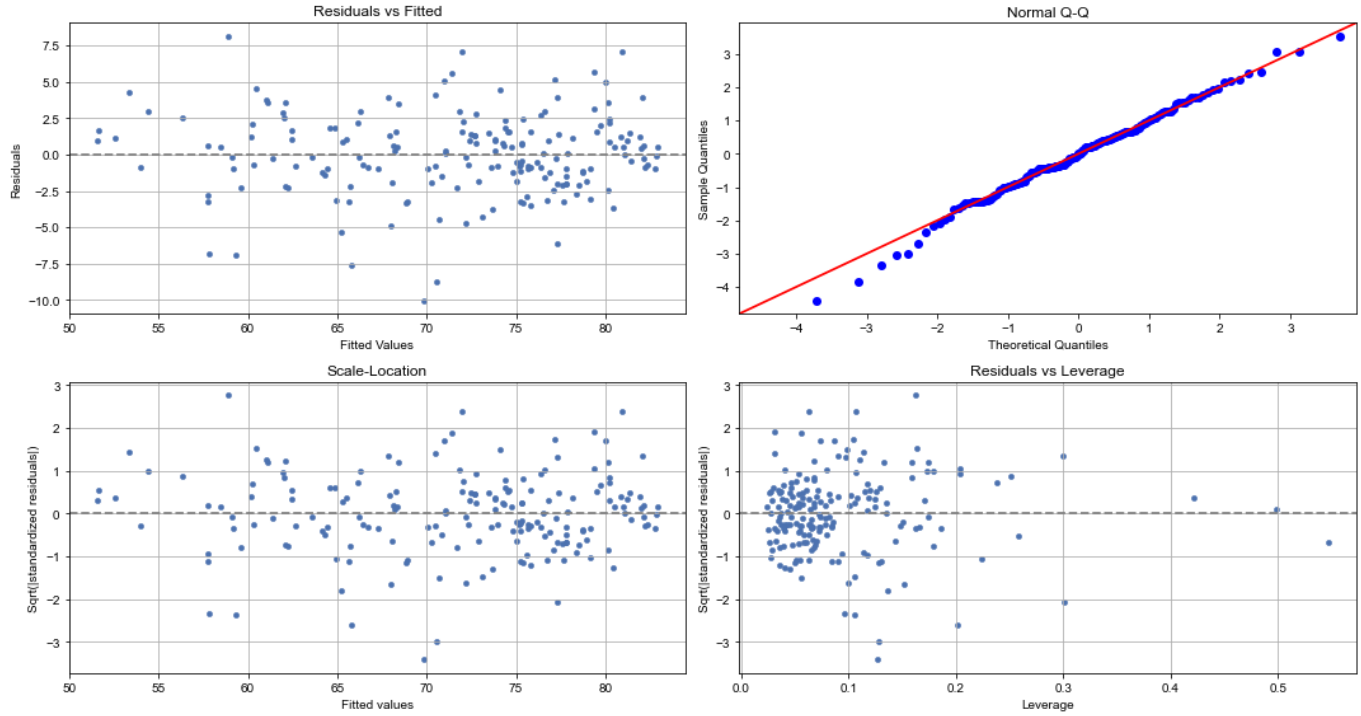
Dep. Variable:	lifeexpectancy	**R-squared:**	0.880
Model:	OLS	**Adj. R-squared:**	0.869

Method:	Least Squares	**F-statistic:**	76.43
Date:	Thu, 09 Dec 2021	**Prob (F-statistic):**	8.29e-68
Time:	19:55:55	**Log-Likelihood:**	-448.14
No. Observations:	183	**AIC:**	930.3
Df Residuals:	166	**BIC:**	984.8
Df Model:	16		
Covariance Type:	nonrobust		

```
plot_ccpr(model, l.columns[1:])
```

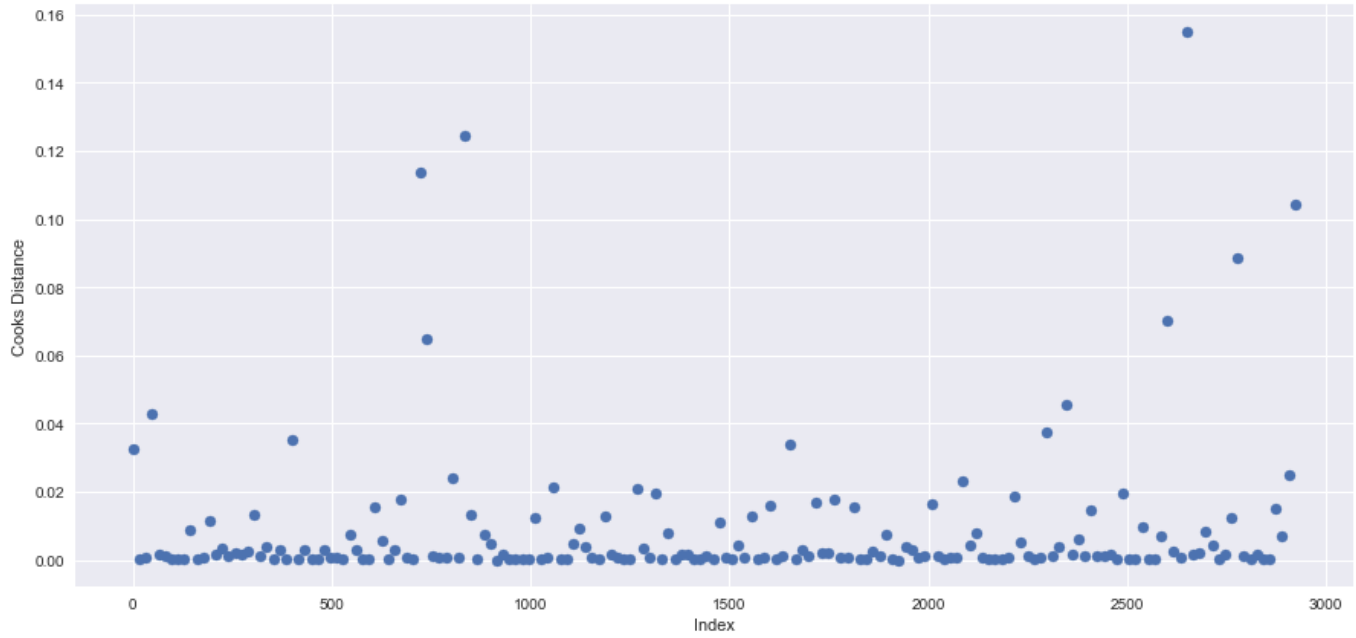



```
plot_model(1, model)
```

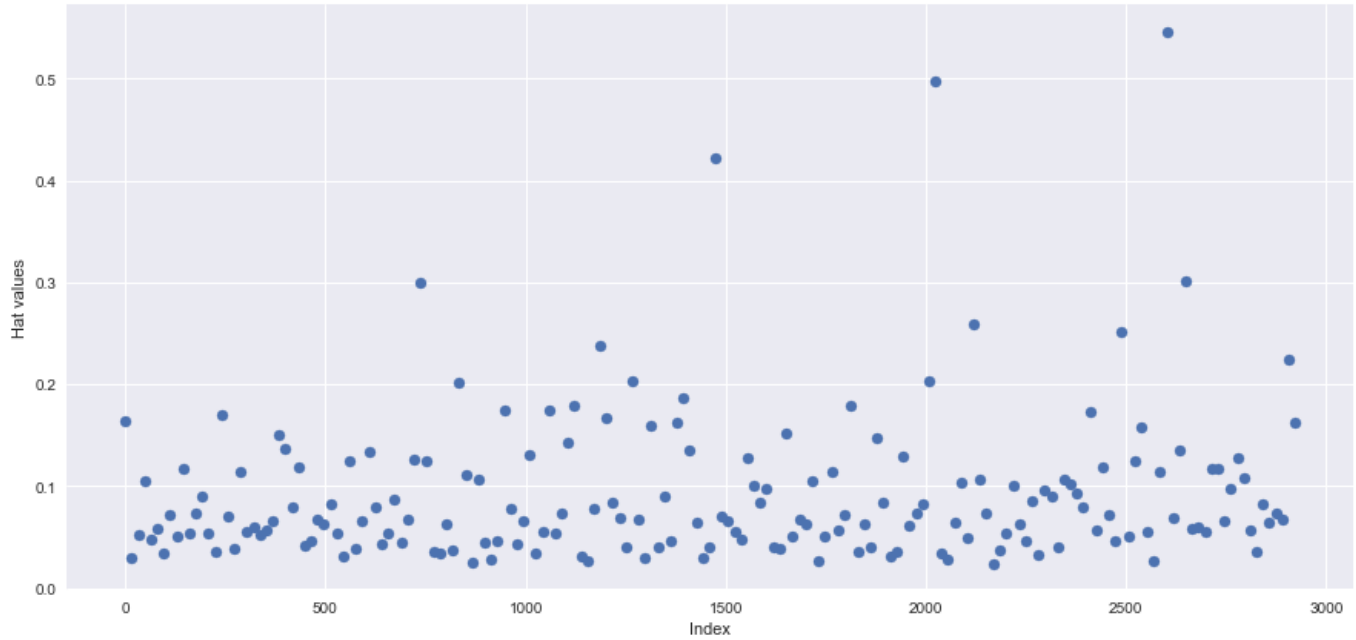



```
influence = model.get_influence()
df = influence.summary_frame()
df.columns
```

```
Index(['dfb_Intercept', 'dfb_adultmortality', 'dfb_infantdeaths',
       'dfb_alcohol', 'dfb_hepatitisb', 'dfb_measles', 'dfb_bmi',
       'dfb_underfivedeaths', 'dfb_polio', 'dfb_totalexpenditure',
       'dfb_diphtheria', 'dfb_hivaids', 'dfb_thinness119years',
       'dfb_thinness59years', 'dfb_incomecompositionofresources',
       'dfb_schooling', 'dfb_gdppercapita', 'cooks_d', 'standard_resid',
       'hat_diag', 'dffits_internal', 'student_resid', 'dffits'],
      dtype='object')
plt.figure(figsize=(15, 7))
plt.scatter(df.index, df.cooks_d)
plt.xlabel('Index')
plt.ylabel('Cooks Distance')
plt.show()
```




```
plt.figure(figsize=(15, 7))
plt.scatter(df.index, df.hat_diag)
plt.xlabel('Index')
plt.ylabel('Hat values')
plt.show()
```




```
shapiro(model.resid)
ShapiroResult(statistic=0.9822049140930176, pvalue=0.019718153402209282)

name = ["Lagrange multiplier statistic", "p-value", "f-value", "f p-value"]
test = sms.het_breuschpagan(model.resid, model.model.exog)
lzip(name, test)
[('Lagrange multiplier statistic', 29.71506816864176),
 ('p-value', 0.019537018389447873),
 ('f-value', 2.011246823587582),
```

```
 ('f p-value', 0.015021203443304109)]
table = sm.stats.anova_lm(model, typ=2) # Type 2 ANOVA DataFrame
print(table)
```

	sum_sq	df	F	PR(>F)
adultmortality	354.229421	1.0	40.961879	1.524555e-09
infantdeaths	8.493416	1.0	0.982150	3.231111e-01
alcohol	0.360281	1.0	0.041662	8.385161e-01
hepatitisb	0.000181	1.0	0.000021	9.963541e-01
measles	0.167316	1.0	0.019348	8.895423e-01
bmi	0.177549	1.0	0.020531	8.862374e-01
underfivedeaths	12.052665	1.0	1.393729	2.394652e-01
polio	9.165519	1.0	1.059869	3.047427e-01
totalexpenditure	0.534972	1.0	0.061862	8.038838e-01
diphtheria	11.597112	1.0	1.341050	2.485122e-01
hivaids	55.870086	1.0	6.460626	1.194461e-02
thinness119years	2.804713	1.0	0.324327	5.697884e-01
thinness59years	0.376574	1.0	0.043546	8.349569e-01
incomecompositionofresources	358.288808	1.0	41.431293	1.257774e-09
schooling	6.709093	1.0	0.775817	3.796971e-01
gdppercapita	15.306888	1.0	1.770036	1.852024e-01
Residual	1435.531881	166.0	NaN	NaN