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).

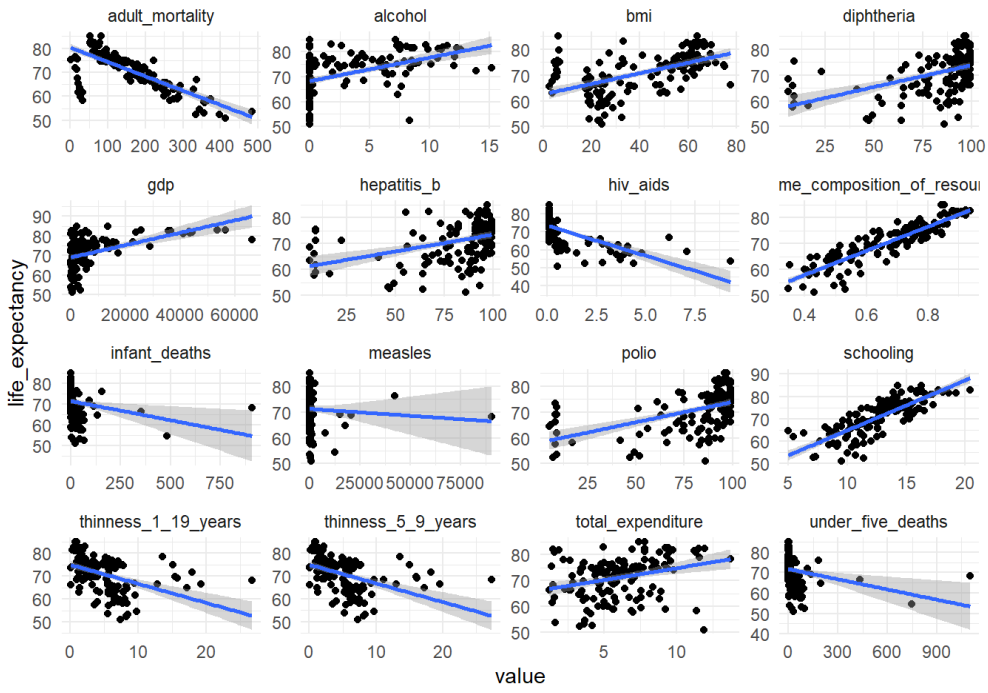
Atliktos analizės aprašymas

1. Naudojant R

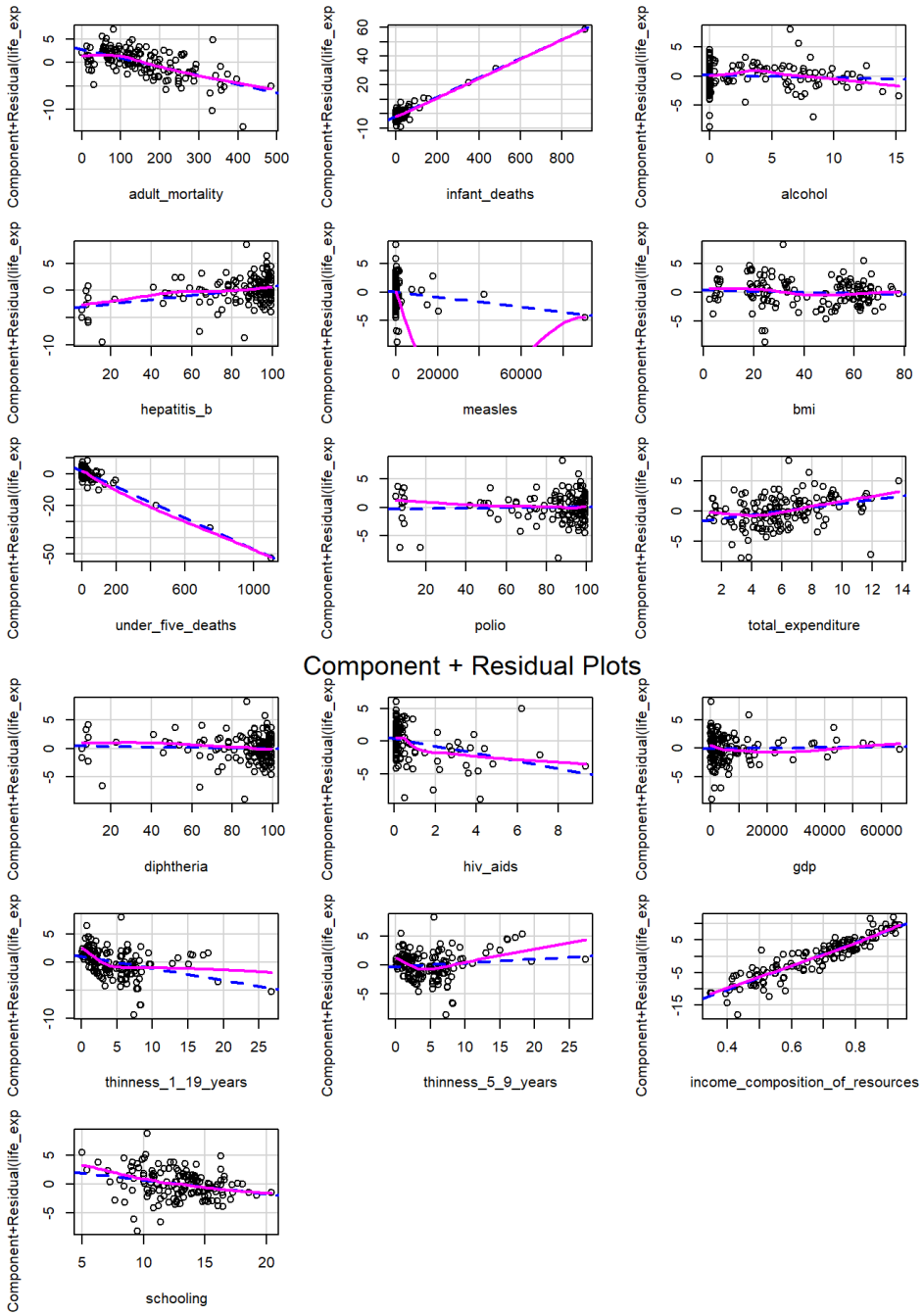
```
library(tidyverse)
library(car)
library(janitor)
x <- read_csv("life.csv") %>% clean_names()
```

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

```
set.seed(150)
transform_1<- function(x) {
  x %>%
  group_by(country) %>%
  fill(everything(), .direction = "up") %>%
  dplyr::select(-c(1, 3), -population, -percentage_expenditure) %>%
  drop_na() %>%
  ungroup()
}

x <- transform_1(x)

x_1 <- x %>% filter(year == max(year)) %>% select(-2)
countries <- x_1$country
x_1 <- x_1 %>% select(-1)

# atskiri duomenys, patikrinti kaip gautas galutinis modelis prognozuoja reikšmes
x_predict <- x %>% filter(year != max(year)) %>% slice_sample(n=10) %>% select(-c(1,2))


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



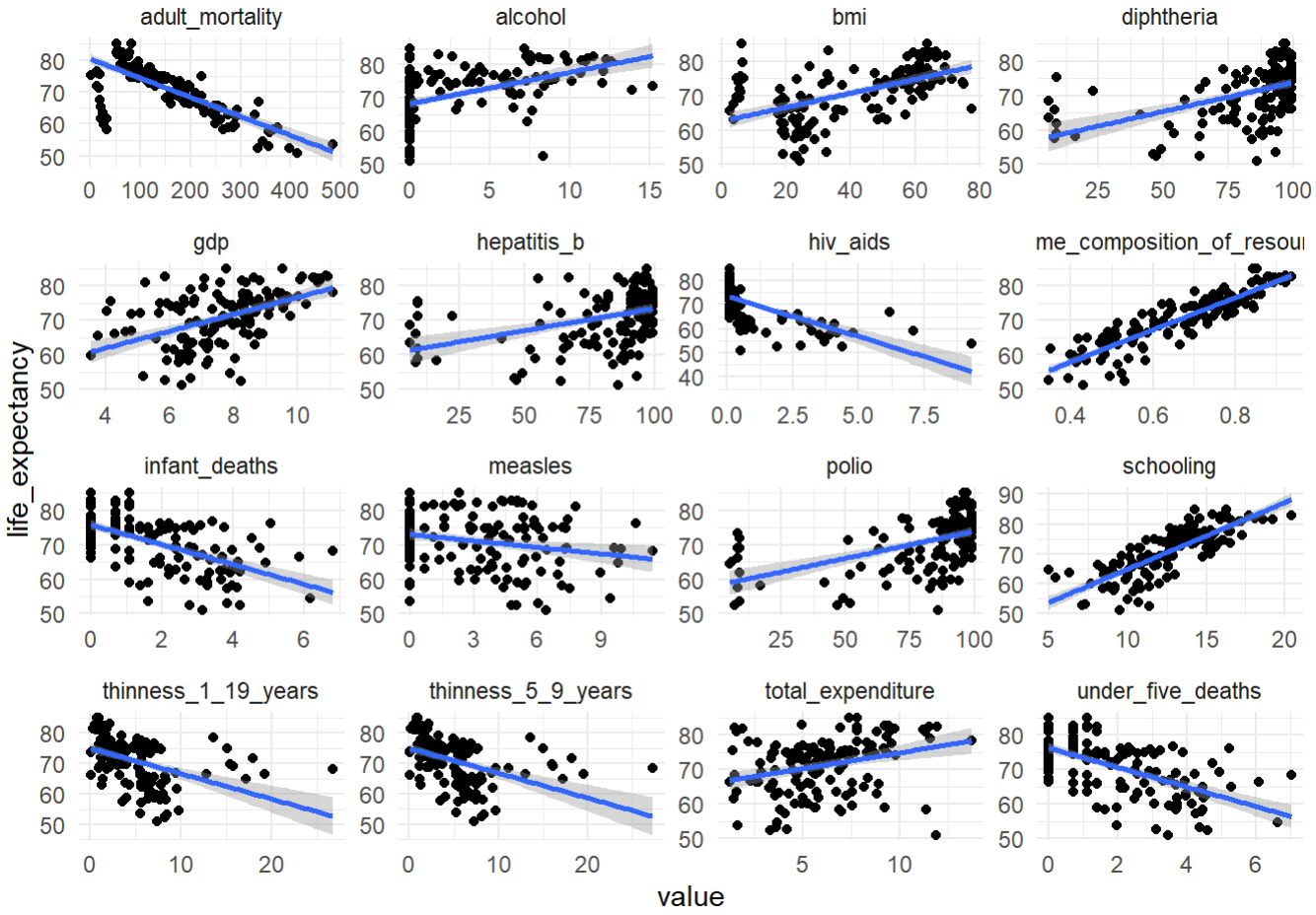

```
model <- lm(life_expectancy ~ ., data = x_1)
crPlots(model)
```



Rasta netiesinė priklausomybė tarp kai kurių kovariančių ir priklausomojo kintamojo. Kintamiesiems “gdp”, “infant_deaths”, “measles” 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),
    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),scales="fre
e") + geom_point() + geom_smooth(method="lm") + theme_minimal()
```

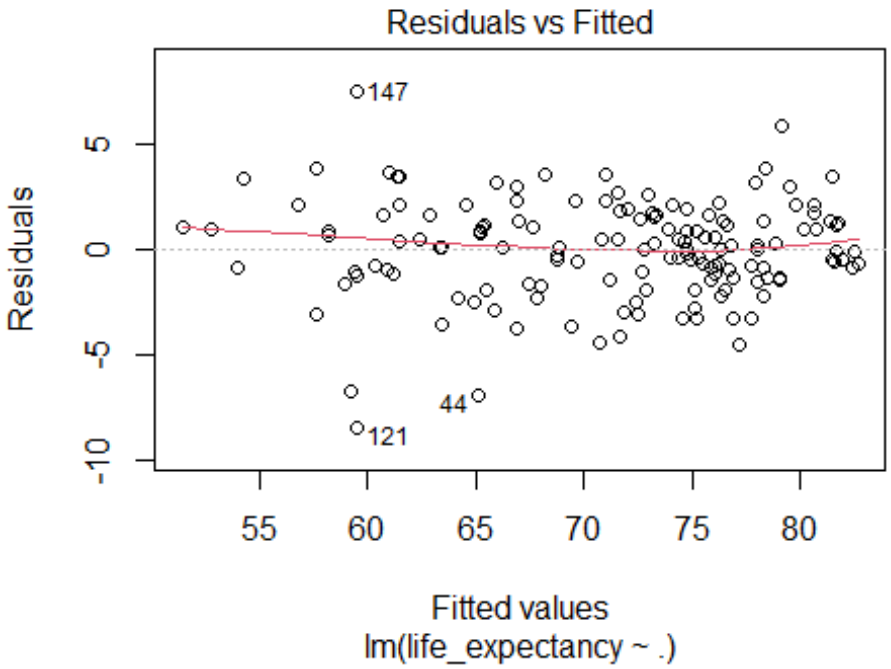


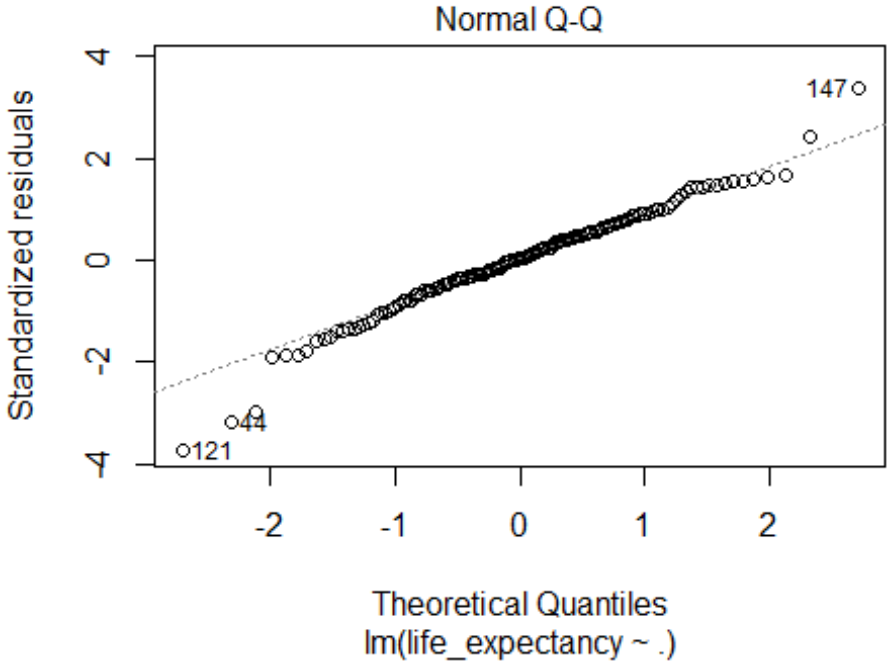
Modifikuoti duomenys išsaugomi faile „life_modified.csv“.

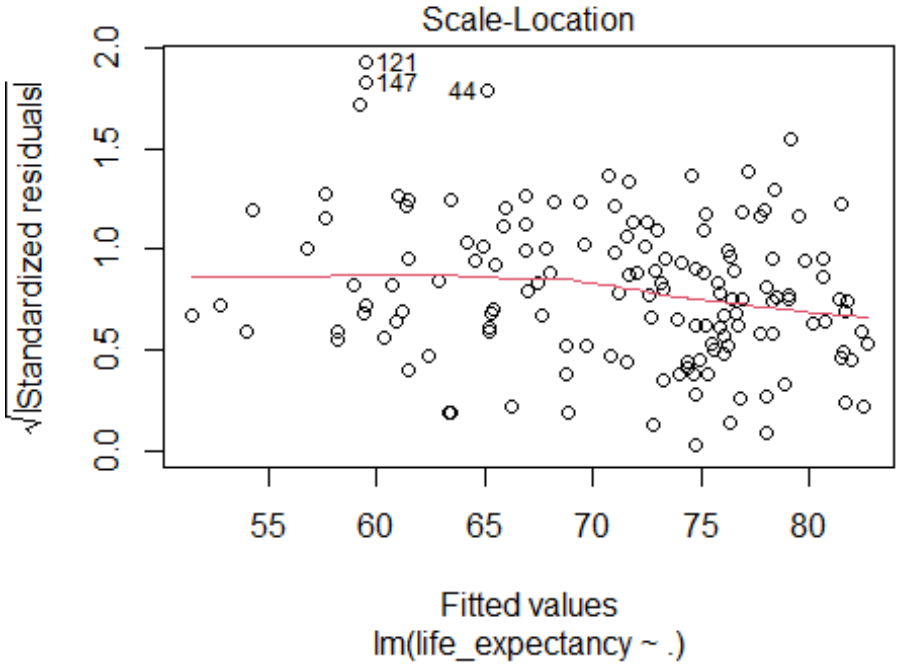
```
Write_csv(x_2, "life_modified.csv")

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

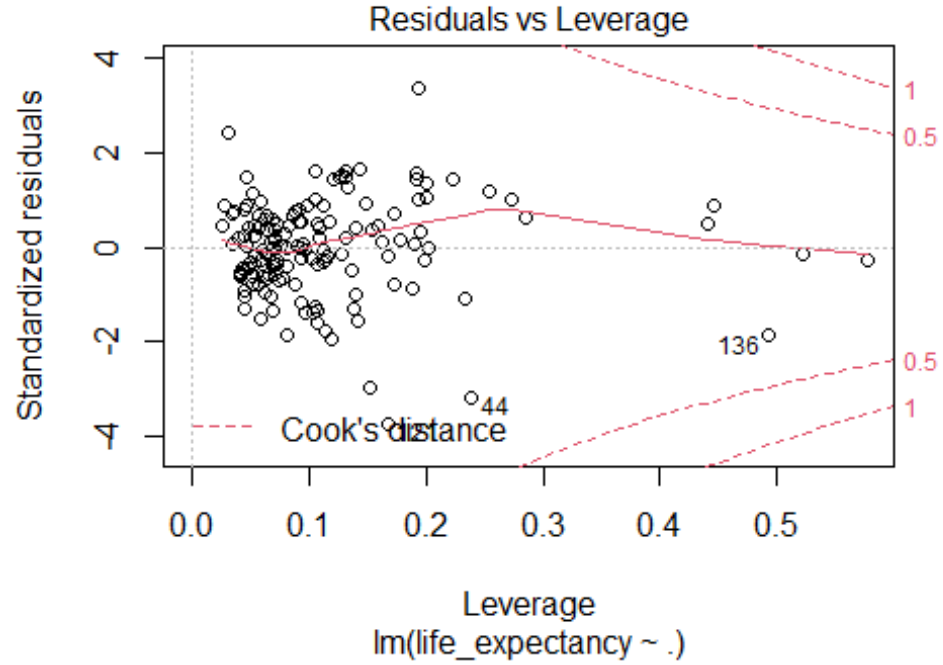
Modelio prielaidos

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

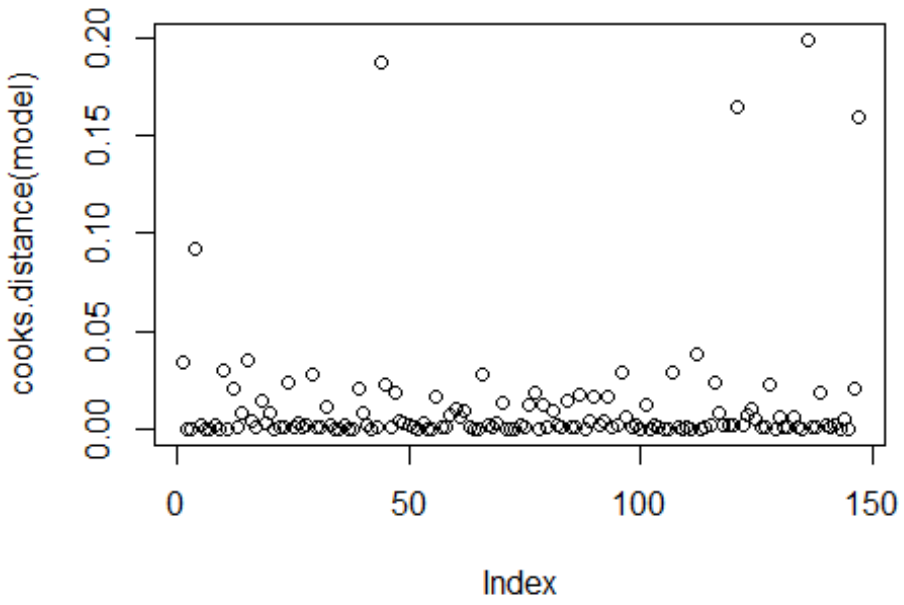

Residuals vs Fitted
Residuals
Fitted values
lm(life_expectancy ~ .)
147
44
121
5
0
-5
-10
55
60
65
70
75
80


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(life_expectancy ~ .)
147
44
121
4
2
0
-2
-4
-2
-1
0
1
2

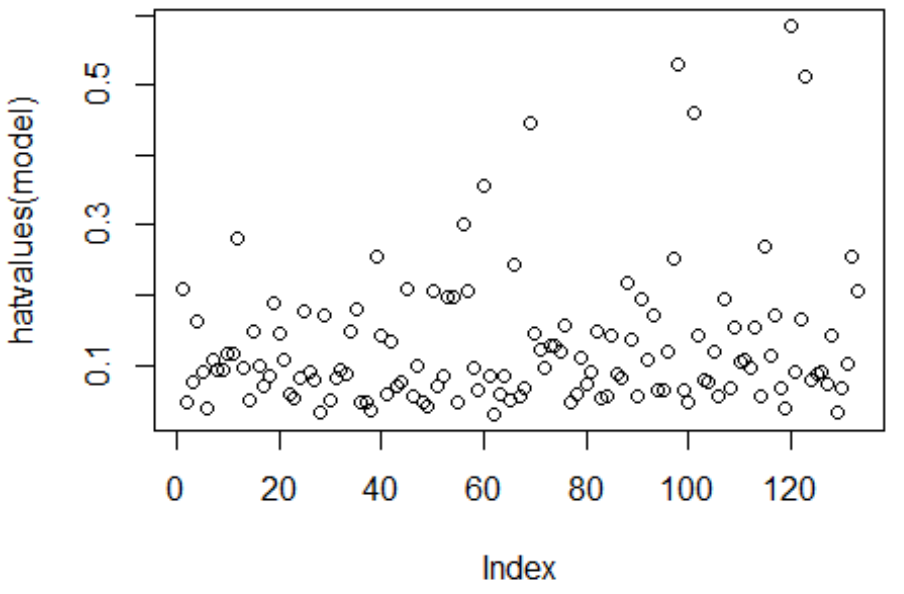





```
plot(cooks.distance(model))
```




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




```
outliers <- c(121,147,44,4)

# patikrinu pagal kokį kintamąjį išsiskiria šios reikšmės
for (i in outliers) {
  for (j in names(x_2)) {
    val <- ecdf(x_2[[j]])(x_2[i,j])
    if (val > 0.95 || val < 0.05) {
```

```
      print(paste(i,countries[i],j,val))
    }
  }
}
```

```
## [1] "121 Sierra Leone life_expectancy 0.00680272108843537"
## [1] "121 Sierra Leone adult_mortality 0.993197278911565"
## [1] "121 Sierra Leone total_expenditure 0.986394557823129"
## [1] "147 Zimbabwe hiv_aids 0.986394557823129"
## [1] "147 Zimbabwe gdp 0.0476190476190476"
## [1] "44 Equatorial Guinea hiv_aids 0.965986394557823"
## [1] "4 Angola life_expectancy 0.0136054421768707"
## [1] "4 Angola infant_deaths 0.952380952380952"
## [1] "4 Angola under_five_deaths 0.952380952380952"
## [1] "4 Angola polio 0.0204081632653061"
```

```
x_3 <- x_2[-outliers,]
write_csv(x_3,"life_modified_no_outliers.csv")

model <- lm(life_expectancy ~ ., data = x_3)
model_outliers <- lm(life_expectancy ~ ., data=x_2)


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

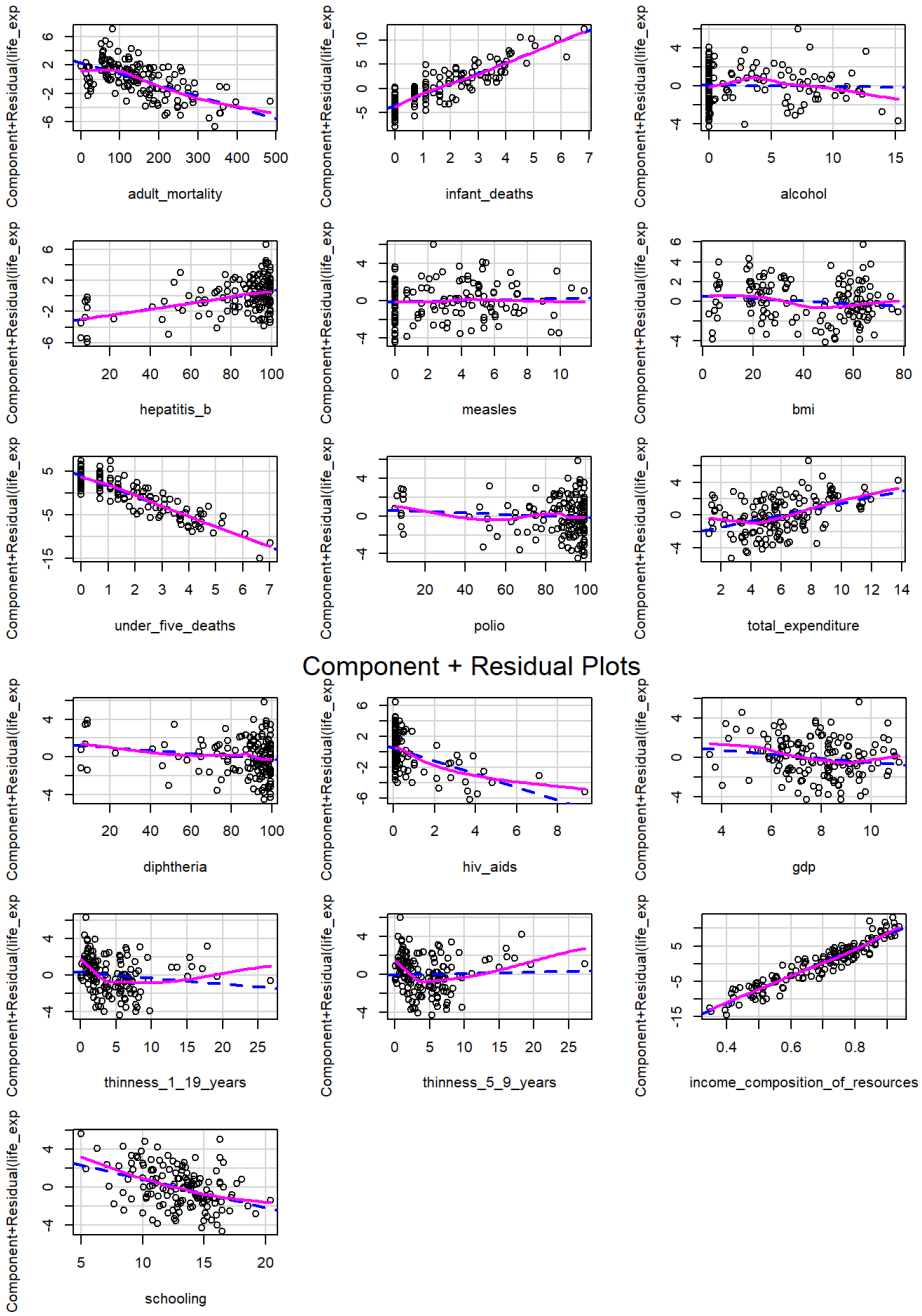
```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model)
## W = 0.9936, p-value = 0.7765
```

```
# Homoskedastiškumo testas
library(lmtest)
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 11.839, df = 16, p-value = 0.755
```

```
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 4630.6  4630.6 1065.6958 < 2.2e-16 ***
## infant_deaths             1  696.6   696.6  160.3292 < 2.2e-16 ***
## alcohol                   1  522.7   522.7  120.2927 < 2.2e-16 ***
## hepatitis_b               1  178.3   178.3   41.0412 2.700e-09 ***
## measles                   1   15.5    15.5    3.5773 0.0608683 .
## bmi                       1  119.4   119.4   27.4788 6.491e-07 ***
## under_five_deaths         1  222.1   222.1   51.1060 6.272e-11 ***
## polio                     1   35.0    35.0    8.0573 0.0052846 **
## total_expenditure         1   64.3    64.3   14.7886 0.0001899 ***
```

```
## diphtheria                          1    7.1     7.1    1.6262 0.2045783
## hiv_aids                            1   79.0    79.0   18.1924 3.885e-05 ***
## gdp                                 1   65.8    65.8   15.1471 0.0001603 ***
## thinness_1_19_years                 1   48.8    48.8   11.2285 0.0010634 **
## thinness_5_9_years                  1    2.0     2.0    0.4700 0.4942645
## income_composition_of_resources     1  791.8   791.8  182.2314 < 2.2e-16 ***
## schooling                           1   13.9    13.9    3.2045 0.0758373 .
## Residuals                         126  547.5     4.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Hipotezė apie reikšmingų kovariančių nebuvimą atmetama.

Modelio parinkimas

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

```
# Požinksninė regresija
library(RcmdrMisc)
model_2 <- stepwise(model,direction = "forward/backward")

##
## Direction:  forward/backward
## Criterion:  BIC
##
## Step:  AIC=293.72
## life_expectancy ~ adult_mortality + hepatitis_b + total_expenditure +
##     hiv_aids + income_composition_of_resources
##
##                                   Df Sum of Sq     RSS    AIC
## <none>                                         884.32 293.72
## + gdp                              1     25.38  858.94 294.43
## + measles                          1     12.07  872.24 296.69
## + thinness_1_19_years              1     10.52  873.80 296.95
## + schooling                        1     10.02  874.29 297.03
## + thinness_5_9_years               1      6.90  877.42 297.56
## + under_five_deaths                1      5.30  879.02 297.82
## + infant_deaths                    1      3.24  881.08 298.17
## + bmi                              1      1.91  882.41 298.39
## + polio                            1      1.80  882.52 298.41
## + alcohol                          1      1.74  882.57 298.42
## + diphtheria                       1      0.08  884.23 298.69
## - hiv_aids                         1     72.84  957.16 300.36
## - total_expenditure                1     75.91  960.22 300.83
## - hepatitis_b                      1     76.85  961.16 300.98
## - adult_mortality                  1    240.59 1124.90 324.10
## - income_composition_of_resources  1   2044.89 2929.20 464.78
```

Parametrų vertinimas ir interpretacija

```
# Pastebimas stiprus koeficientų reikšmių skirtumas tarp modelio su išskirtimis ir be
(coef(model_2) - coef(model_outliers_2)) / coef(model_2)

##                     (Intercept) income_composition_of_resources
##                    1.658413e-02                    1.000591e+00
##                  adult_mortality                        hiv_aids
##                    2.996389e+00                    1.313718e+00
##                total_expenditure                      hepatitis_b
##                    2.986259e+00                   -1.406753e+03
```

```r
# Koeficientai
summary(model_2)
```

```
##
## Call:
## lm(formula = life_expectancy ~ income_composition_of_resources +
##     adult_mortality + hiv_aids + total_expenditure + hepatitis_b,
##     data = x_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1803 -1.1947 -0.0956  1.4552  5.8049
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    48.311111   1.380377  34.998  < 2e-16 ***
## income_composition_of_resources 32.502820   1.547454  21.004  < 2e-16 ***
## adult_mortality                -0.016440   0.002873  -5.722 6.36e-08 ***
## hiv_aids                       -0.923119   0.183090  -5.042 1.43e-06 ***
## total_expenditure               0.330922   0.072742   4.549 1.17e-05 ***
## hepatitis_b                     0.023522   0.008026   2.931  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.103 on 137 degrees of freedom
## Multiple R-squared:  0.9247, Adjusted R-squared:  0.9219
## F-statistic: 336.3 on 5 and 137 DF,  p-value: < 2.2e-16
```

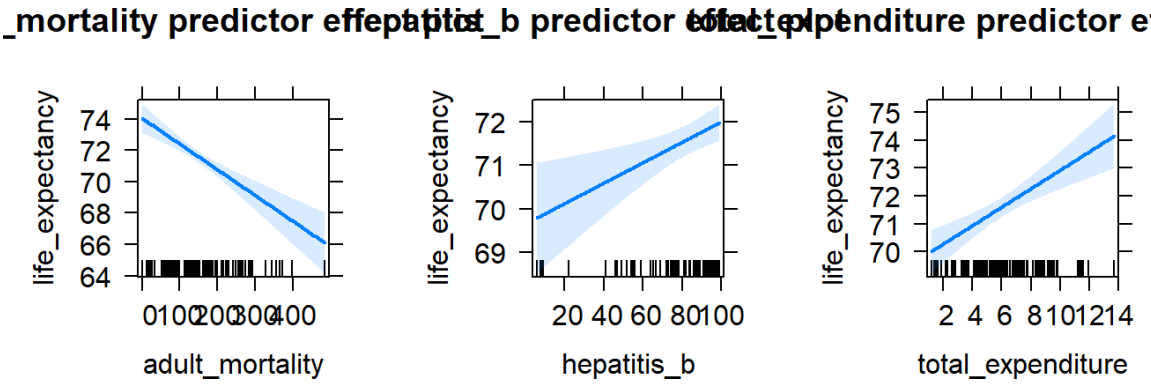
```r
  # Visų koeficientų interpretacija paprasta,
  # nes pažingsnine regresija neišrinkti transformuoti kintamieji
library(lm.beta)
# Standartizuoti koeficientai
lm.beta(model_2)
```

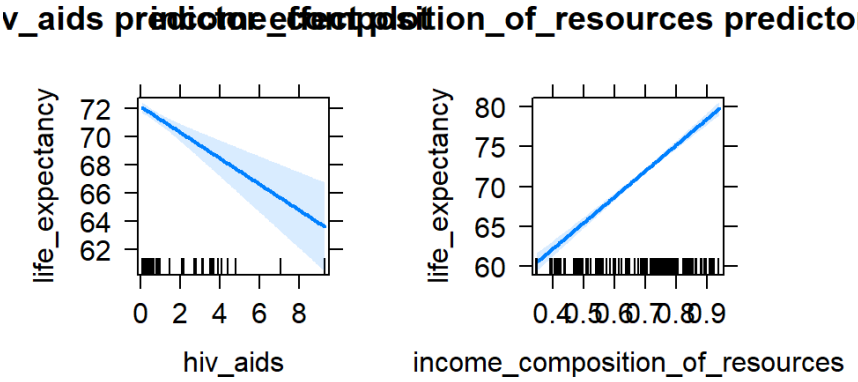
```
##
## Call:
## lm(formula = life_expectancy ~ income_composition_of_resources +
##     adult_mortality + hiv_aids + total_expenditure + hepatitis_b,
##     data = x_3)
##
## Standardized Coefficients::
##                     (Intercept) income_composition_of_resources
##                      0.00000000                       0.65293000
##                  adult_mortality                         hiv_aids
##                     -0.20565299                      -0.16600144
##                total_expenditure                      hepatitis_b
##                      0.11157918                       0.07489474
```

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

```
##                                       2.5 %      97.5 %
## (Intercept)                    45.581510927 51.04071071
## income_composition_of_resources 29.442835792 35.56280374
## adult_mortality                -0.022122297 -0.01075856
## hiv_aids                       -1.285166225 -0.56107189
## total_expenditure               0.187080539  0.47476436
## hepatitis_b                     0.007652056  0.03939220
```

```r
# Kovariancių 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 – imunizacija nuo hepatito B (stulp. hepatitis_b 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  0.25778500        0.09263737
## hepatitis_b                         0.25778500  1.00000000       -0.03602603
## total_expenditure                   0.09263737 -0.03602603        1.00000000
## hiv_aids                            0.29146768 -0.18202475        0.13459109
## income_composition_of_resources     0.15257421 -0.12052089       -0.14829262
## life_expectancy                    -0.46246417  0.28275567        0.28115852
##                                 hiv_aids income_composition_of_resources
## adult_mortality                 0.2914677                       0.1525742
## hepatitis_b                    -0.1820247                      -0.1205209
## total_expenditure               0.1345911                      -0.1482926
## hiv_aids                        1.0000000                       0.1911298
## income_composition_of_resources 0.1911298                       1.0000000
## life_expectancy                -0.2758631                       0.8355260
##                                life_expectancy
## adult_mortality                     -0.4624642
## hepatitis_b                          0.2827557
## total_expenditure                    0.2811585
## hiv_aids                            -0.2758631
## income_composition_of_resources      0.8355260
## life_expectancy                      1.0000000

# Variance inflation factor
vif(model_2)

## income_composition_of_resources                adult_mortality
##                        1.757174                       2.349157
##                        hiv_aids               total_expenditure
##                        1.971169                       1.093881
##                      hepatitis_b
##                        1.187387
```

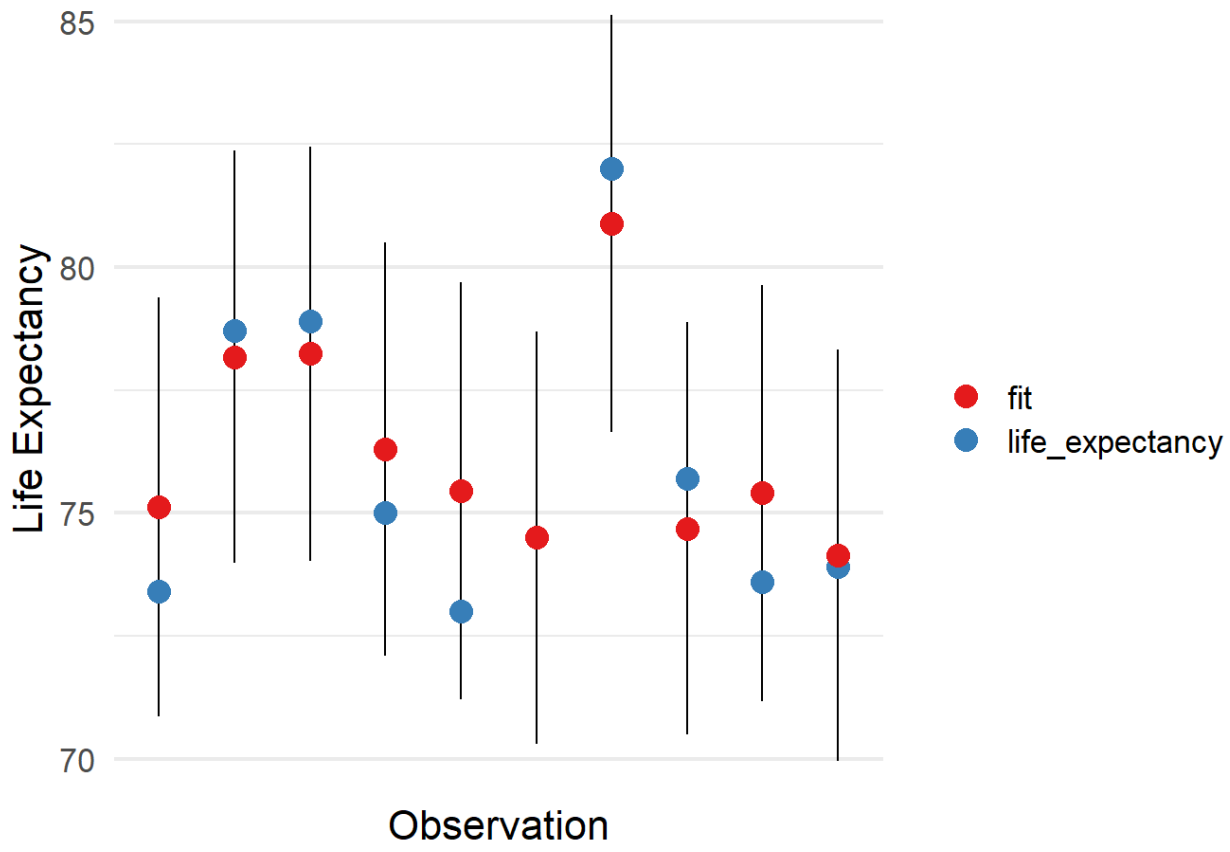
Naudojant dalinių koreliacijų matricą nerasta stiprių kovariančių tarpusavio koreliacijų. Variance inflation factor reiškmės <2.35 visoms modelyje esančioms kovariantėms. Pasirinkus VIF ribą 4 tariame, kad reiškmingo multikolinearumo modelyje nėra.

Modelio tinkamumo analizė

```
summary(model_2)

##
## Call:
## lm(formula = life_expectancy ~ income_composition_of_resources +
##     adult_mortality + hiv_aids + total_expenditure + hepatitis_b,
##     data = x_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1803 -1.1947 -0.0956  1.4552  5.8049
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    48.311111   1.380377  34.998  < 2e-16 ***
## income_composition_of_resources 32.502820   1.547454  21.004  < 2e-16 ***
## adult_mortality                -0.016440   0.002873  -5.722 6.36e-08 ***
## hiv_aids                       -0.923119   0.183090  -5.042 1.43e-06 ***
## total_expenditure               0.330922   0.072742   4.549 1.17e-05 ***
## hepatitis_b                     0.023522   0.008026   2.931  0.00396 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.103 on 137 degrees of freedom
## Multiple R-squared:  0.9247, Adjusted R-squared:  0.9219
## F-statistic: 336.3 on 5 and 137 DF,  p-value: < 2.2e-16
```

```
  # R-squared = 0.925
  # Adj R-squared = 0.922

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") +
  theme_minimal(base_size = 16) +
  scale_color_brewer("",palette = "Set1")
}

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

Modelis paaiškina 92.5% duomenų sklaidos $R^2 = 0.925$. 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 92.5% duomenų sklaidos ($F(5,137) = 336.3, R^2 = 0.925, p<0.001$). 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* $\beta=-0.21, p<0.001$)
Imunizacija nuo Hepatito B tarp 1 metų vaikų % (stulp. hepatitis_b $\beta=0.07, p=0.003$)
Dalis visų vyriausybės išlaidų sveikatos apsaugai (stulp. total_expenditure $\beta=0.11, p<0.001$)
Mirtys nuo ŽIV/AIDS nuo 0 iki 4 metų 1000 gimimų (stulp. hiv_aids $\beta=-0.17, p<0.001$)
HDI pagal pajamų parametrą (stulp. income_composition_of_resources $\beta=0.65, p<0.001$)

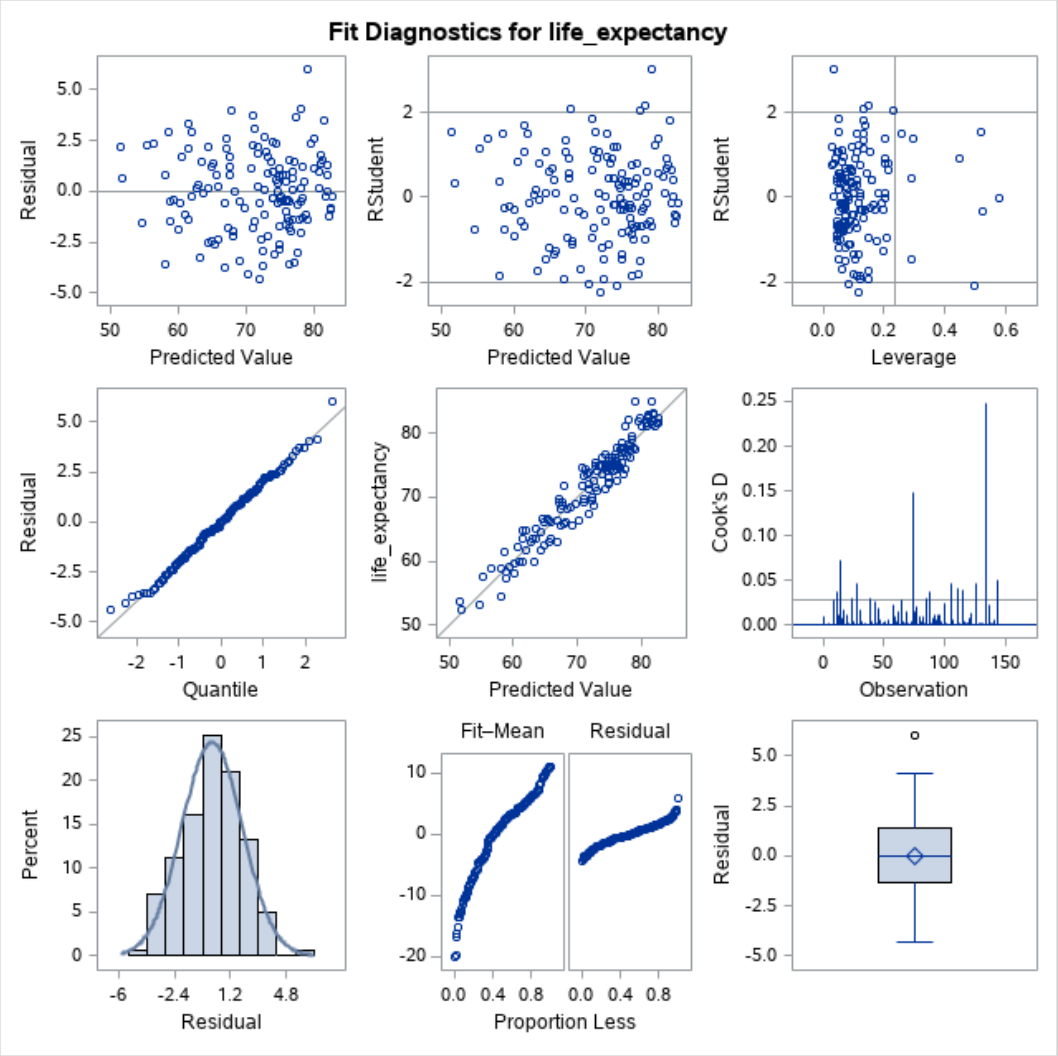
2. Naudojant SAS

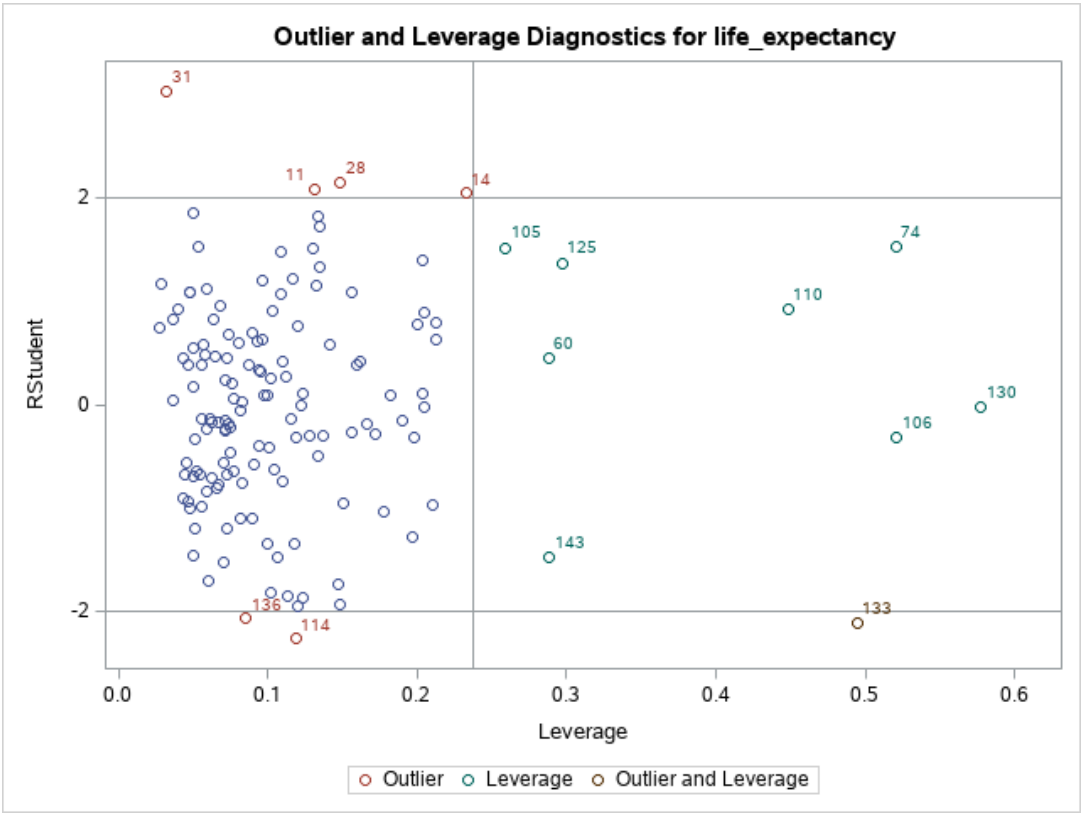
Naudojamas anksčiau sukurtas duomenų failas.

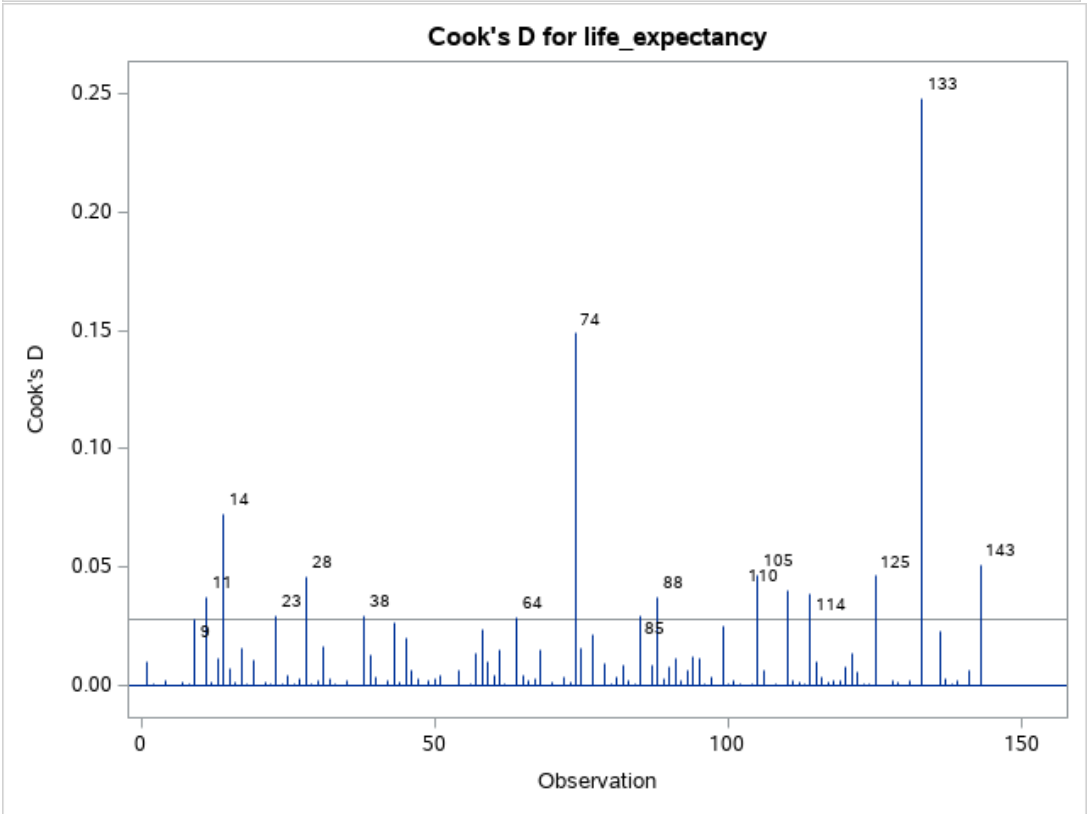
```
PROC IMPORT DATAFILE='/home/u45871880/life_modified_no_outliers.csv'
       DBMS=CSV
       OUT=data;
       GETNAMES=YES;
RUN;
```

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;
```




Outlier and Leverage Diagnostics for life_expectancy
RStudent
Leverage
Outlier
Leverage
Outlier and Leverage
31
11
28
14
105
125
74
110
60
130
106
143
136
114
133


Cook's D for life_expectancy
Cook's D
Observation
133
74
14
28
11
23
38
64
9
88
85
105
110
114
125
143

```
/* Normalumo testas */

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

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.993603	Pr < W	0.7765
Kolmogorov-Smirnov	D	0.037915	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.022271	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.16701	Pr > A-Sq	>0.2500

```
/* 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 5
Variable hepatitis_b Entered: R-Square = 0.9247 and C(p) = 8.4164

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7434.69132	1486.93826	336.28	<.0001
Error	137	605.77861	4.42174		
Corrected Total	142	8040.46993			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	48.31111	1.38038	5416.16894	1224.89	<.0001
adult_mortality	-0.01644	0.00287	144.75699	32.74	<.0001
hepatitis_b	0.02352	0.00803	37.98320	8.59	0.0040
total_expenditure	0.33092	0.07274	91.51193	20.70	<.0001
hiv_aids	-0.92312	0.18309	112.40387	25.42	<.0001
income_composition_of_resources	32.50282	1.54745	1950.74236	441.17	<.0001

Bounds on condition number: 2.3492, 41.794
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	0.8285	0.8285	178.396	681.05	<.0001
2	adult_mortality		2	0.0634	0.8919	63.0759	82.09	<.0001
3	hiv_aids		3	0.0162	0.9081	35.0081	24.58	<.0001
4	total_expenditure		4	0.0118	0.9199	15.1580	20.35	<.0001
5	hepatitis_b		5	0.0047	0.9247	8.4164	8.59	0.0040

Matome, kad pažingsninė regresija išrenka tas pačias kovariantes kaip ir atliekant užduotį su R.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	7434.69132	1486.93826	336.28	<.0001
Error	137	605.77861	4.42174		
Corrected Total	142	8040.46993			

Root MSE	2.10279	R-Square	0.9247
Dependent Mean	71.59161	Adj R-Sq	0.9219
Coeff Var	2.93721		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Squared Partial Corr Type II	Variance Inflation	95% Confidence Limits	
Intercept	1	48.31111	1.38038	35.00	<.0001	0	.	0	45.58151	51.04071
adult_mortality	1	-0.01644	0.00287	-5.72	<.0001	-0.20565	0.19287	2.34916	-0.02212	-0.01076
hepatitis_b	1	0.02352	0.00803	2.93	0.0040	0.07489	0.05900	1.18739	0.00765	0.03939
total_expenditure	1	0.33092	0.07274	4.55	<.0001	0.11158	0.13124	1.09388	0.18708	0.47476
hiv_aids	1	-0.92312	0.18309	-5.04	<.0001	-0.16600	0.15651	1.97117	-1.28517	-0.56107
income_composition_of_resources	1	32.50282	1.54745	21.00	<.0001	0.65293	0.76305	1.75717	29.44284	35.56280

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')
```

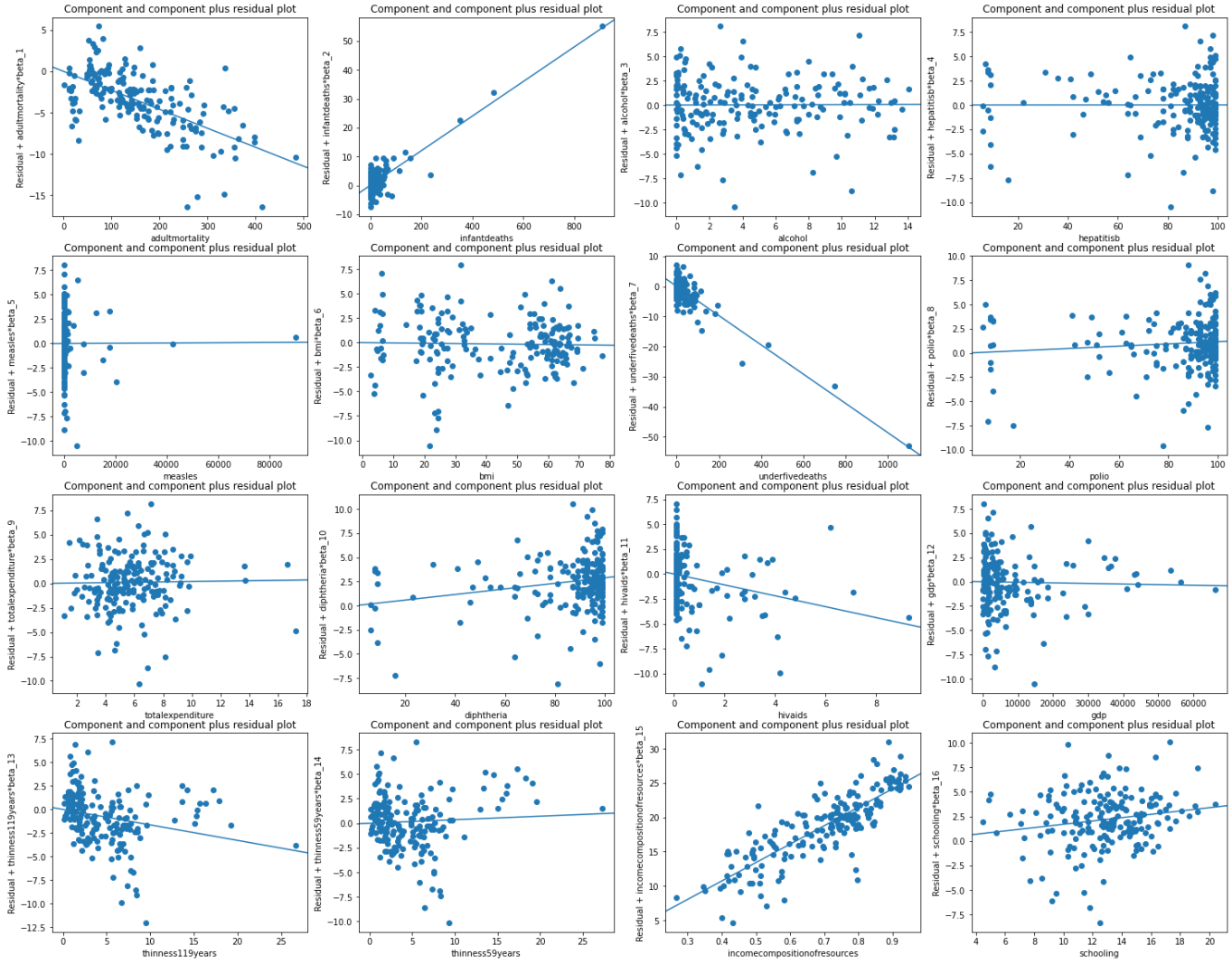
```
    # qqplot
    sm.qqplot(
        df['resid'], dist=stats.t, fit=True, line='45',
        ax=axes[0, 1], c='#4C72B0'
    )
    axes[0, 1].set_title('Normal Q-Q')

    # 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.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", "population", "percentageexpenditure"], axis
= 1)

f = "lifeexpectancy~" + "+".join(d.columns[1:])


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

OLS Regression Results

Dep. Variable:	lifeexpectancy	**R-squared:**	0.882
Model:	OLS	**Adj. R-squared:**	0.871
Method:	Least Squares	**F-statistic:**	77.73
Date:	Mon, 13 Dec 2021	**Prob (F-statistic):**	2.46e-68
Time:	20:08:08	**Log-Likelihood:**	-446.78

No. Observations:	183	**AIC:**	927.6
Df Residuals:	166	**BIC:**	982.1
Df Model:	16		
Covariance Type:	nonrobust		

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



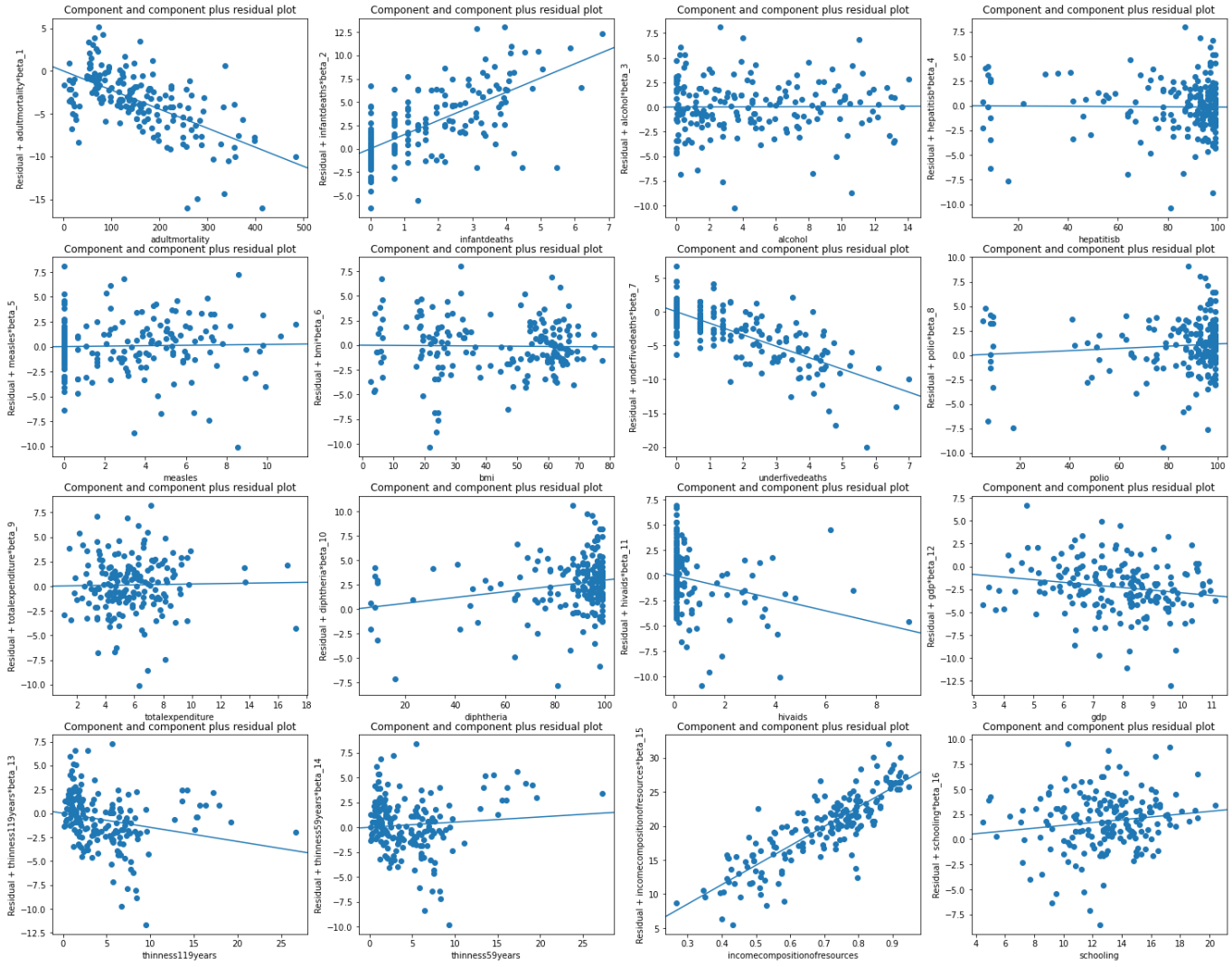
Normalised data

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

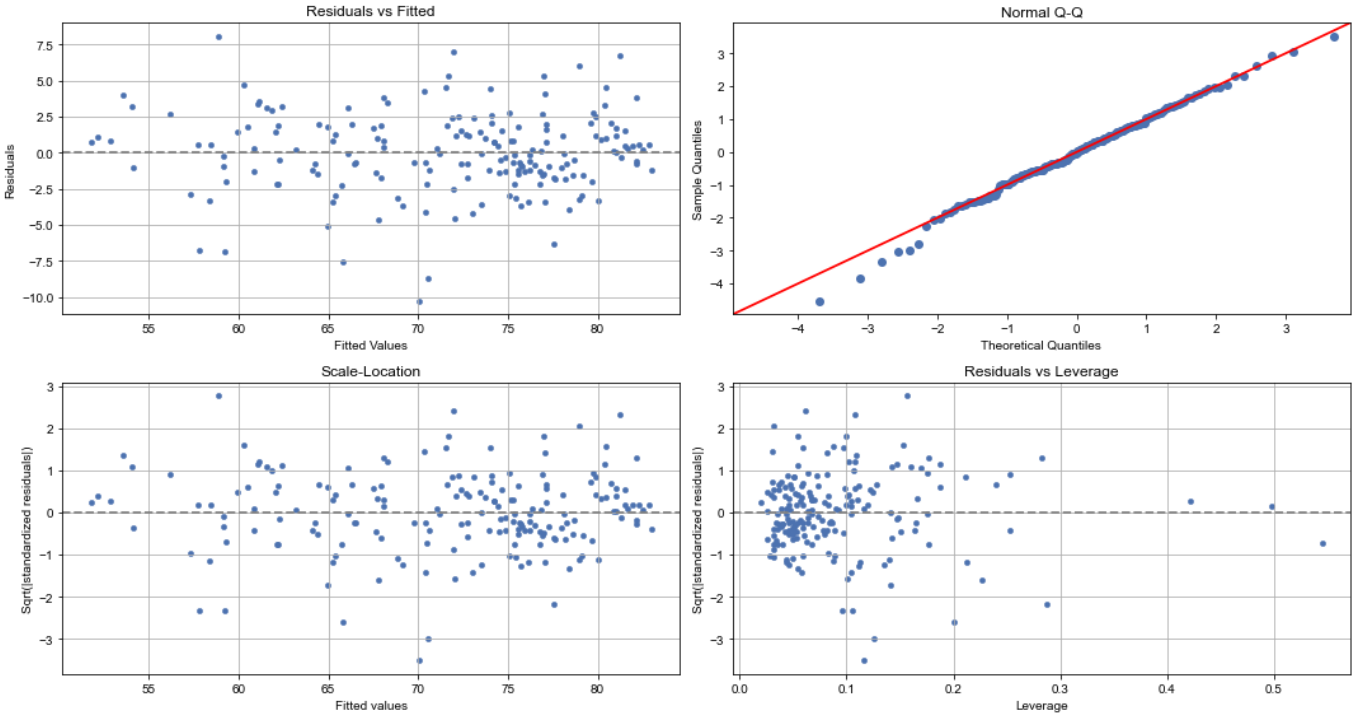
```
model = ols(formula = f, data=l).fit()
model.summary()
```

Dep. Variable:	lifeexpectancy	**R-squared:**	0.881
Model:	OLS	**Adj. R-squared:**	0.870
Method:	Least Squares	**F-statistic:**	77.12
Date:	Mon, 13 Dec 2021	**Prob (F-statistic):**	4.34e-68
Time:	20:08:12	**Log-Likelihood:**	-447.42
No. Observations:	183	**AIC:**	928.8
Df Residuals:	166	**BIC:**	983.4
Df Model:	16		
Covariance Type:	nonrobust		

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



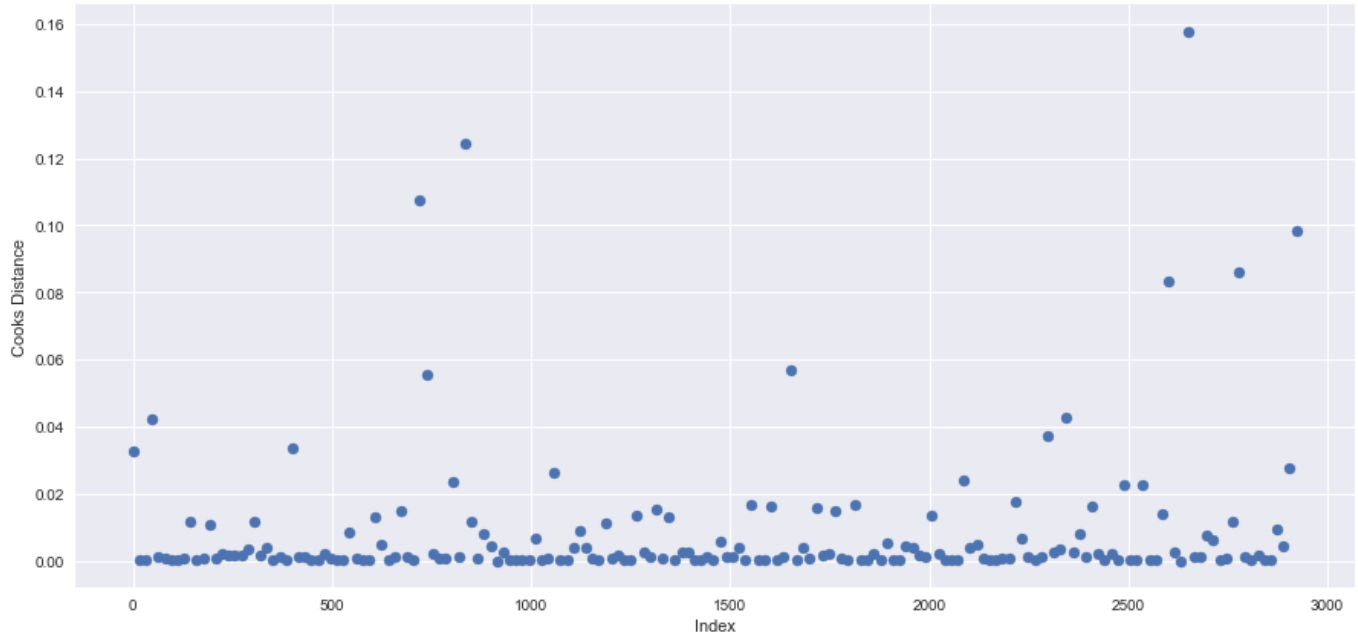
```
plot_model(l, model)
```



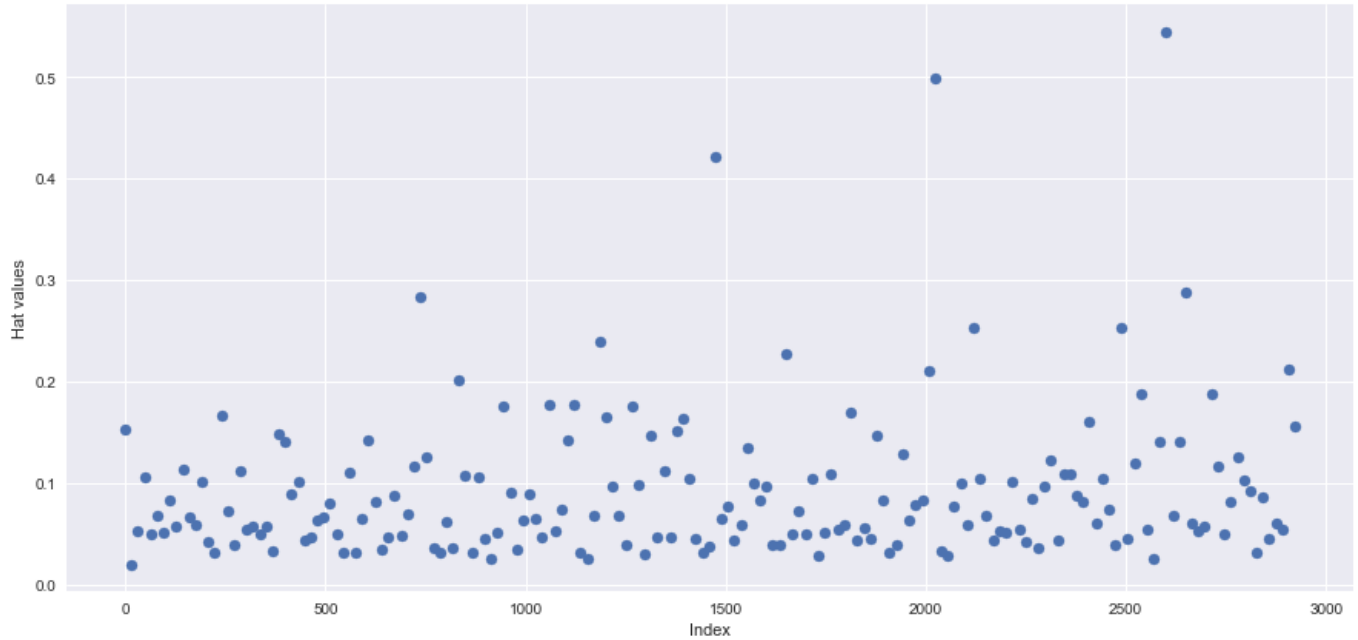

```
influence = model.get_influence()
df = influence.summary_frame()
df.columns
```

```
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_gdp', 'dfb_thinness119years',
       'dfb_thinness59years', 'dfb_incomecompositionofresources',
       'dfb_schooling', '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=0.9821522235870361, pvalue=0.019396508112549782)
```

```
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', 28.793596570070083),
 ('p-value', 0.025365603737385573),
 ('f-value', 1.9372319032796783),
 ('f p-value', 0.020253339084571116)]
```

```
table = sm.stats.anova_lm(model, typ=2) # Type 2 ANOVA DataFrame
print(table)
```

	sum_sq	df	F	PR(>F)
adultmortality	344.576464	1.0	40.160472	2.119607e-09
infantdeaths	7.534654	1.0	0.878166	3.500659e-01
alcohol	0.076618	1.0	0.008930	9.248276e-01
hepatitisb	0.051472	1.0	0.005999	9.383557e-01
measles	0.452394	1.0	0.052727	8.186675e-01
bmi	0.206407	1.0	0.024057	8.769290e-01
underfivedeaths	10.134833	1.0	1.181217	2.786839e-01
polio	6.749995	1.0	0.786714	3.763789e-01
totalexpenditure	0.439789	1.0	0.051257	8.211680e-01
diphtheria	13.729664	1.0	1.600196	2.076484e-01
hivaids	55.264595	1.0	6.441102	1.207151e-02
gdp	26.628496	1.0	3.103558	7.996226e-02
thinness119years	3.112154	1.0	0.362722	5.478200e-01
thinness59years	0.409867	1.0	0.047770	8.272583e-01
incomecompositionofresources	367.860671	1.0	42.874252	6.984015e-10
schooling	4.859405	1.0	0.566365	4.527729e-01
Residual	1424.278414	166.0	NaN	NaN