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).

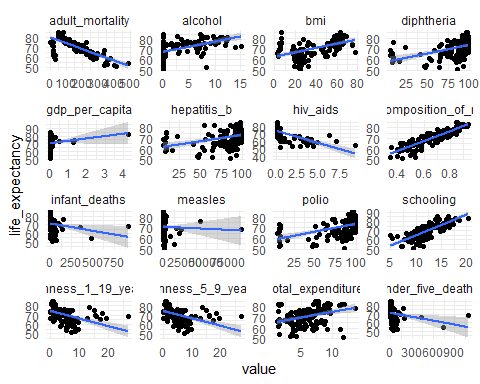
# Atliktos analizės aprašymas

**1. Naudojant R**

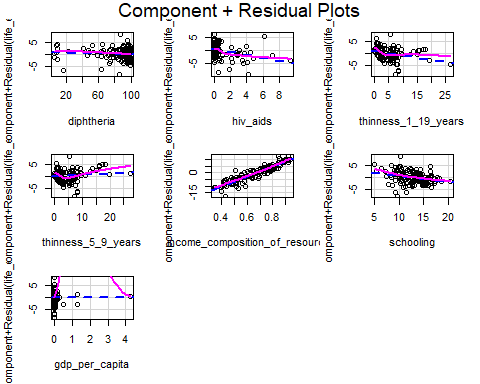
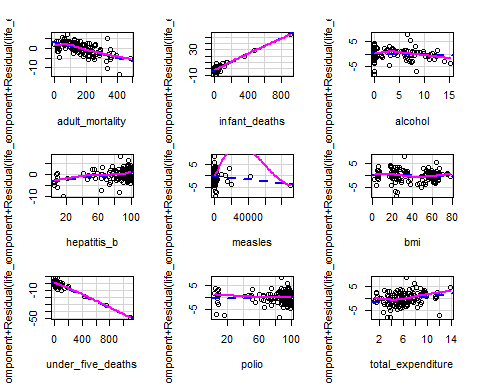
library(tidyverse)  
library(car)  
library(janitor)  
x <- read\_csv("life.csv") %>% clean\_names()

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),scales="free") + geom\_point() + geom\_smooth(method="lm") + theme\_minimal()

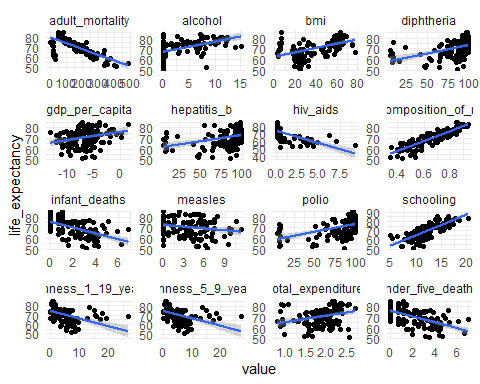


model <- lm(life\_expectancy ~ ., data = x\_1)  
crPlots(model)



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),scales="free") + geom\_point() + geom\_smooth(method="lm") + theme\_minimal()

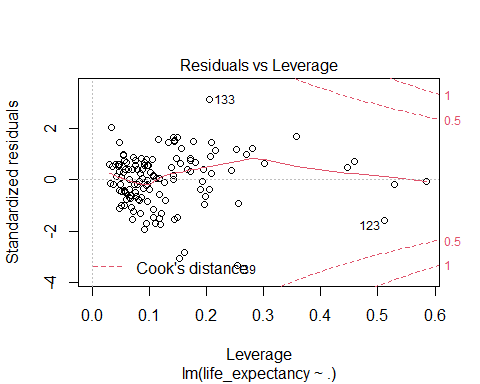
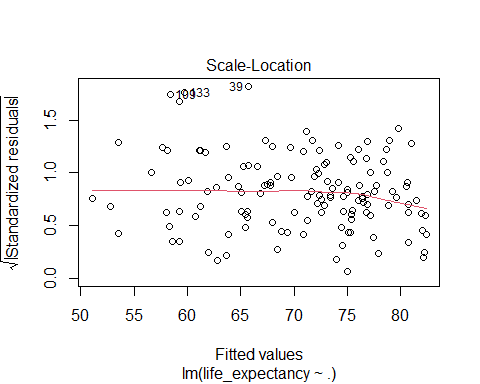
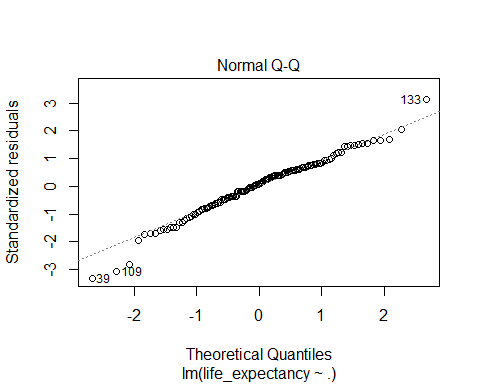
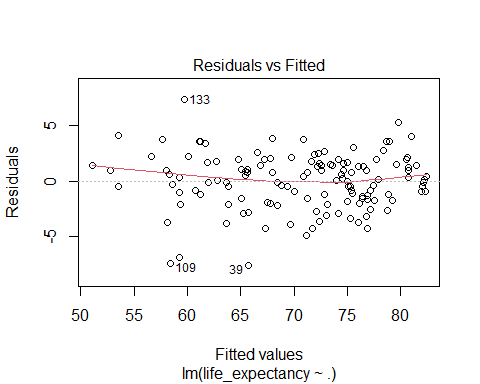


Modifikuoti duomenys išsaugomi faile „life\_modified.csv“.

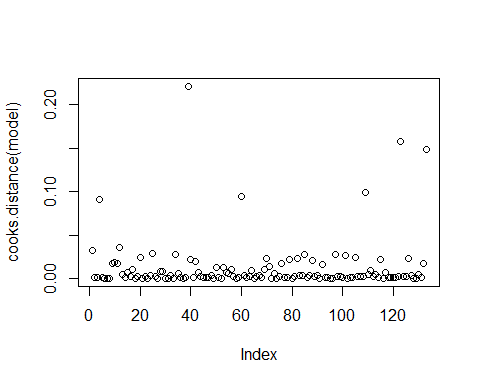
write.csv(x\_2, "life\_modified.csv")  
  
# Sukuriamas modelis  
model <- lm(life\_expectancy ~ ., data = x\_2)

Modelio prielaidos

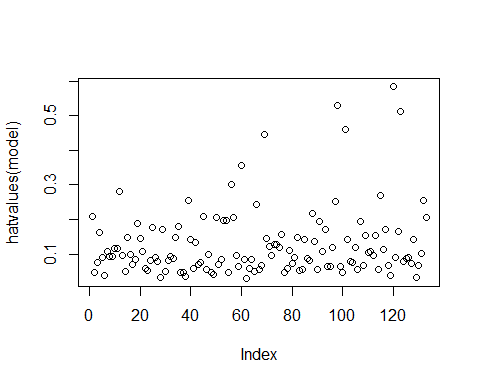
# Tikrinamas liekanų normalumas, homoskadiškumas, liekanų nepriklausomumas, išskirtys  
plot(model)



plot(cooks.distance(model))



plot(hatvalues(model))



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

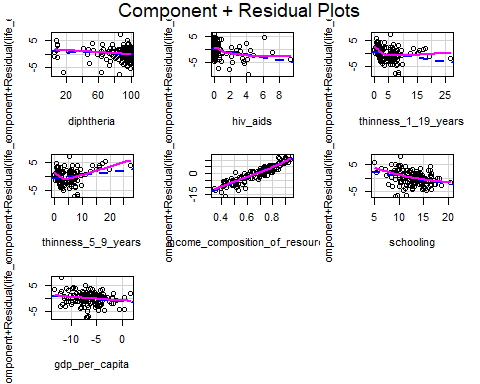
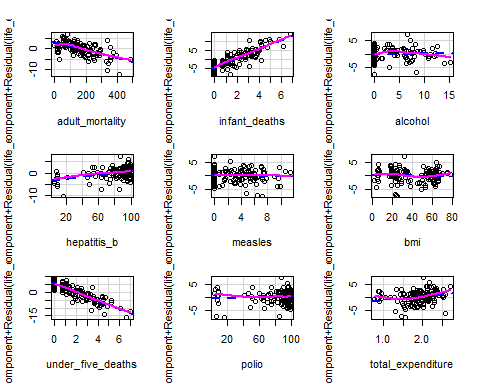
##   
## Shapiro-Wilk normality test  
##   
## W = 0.98195, p-value = 0.07493

# 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

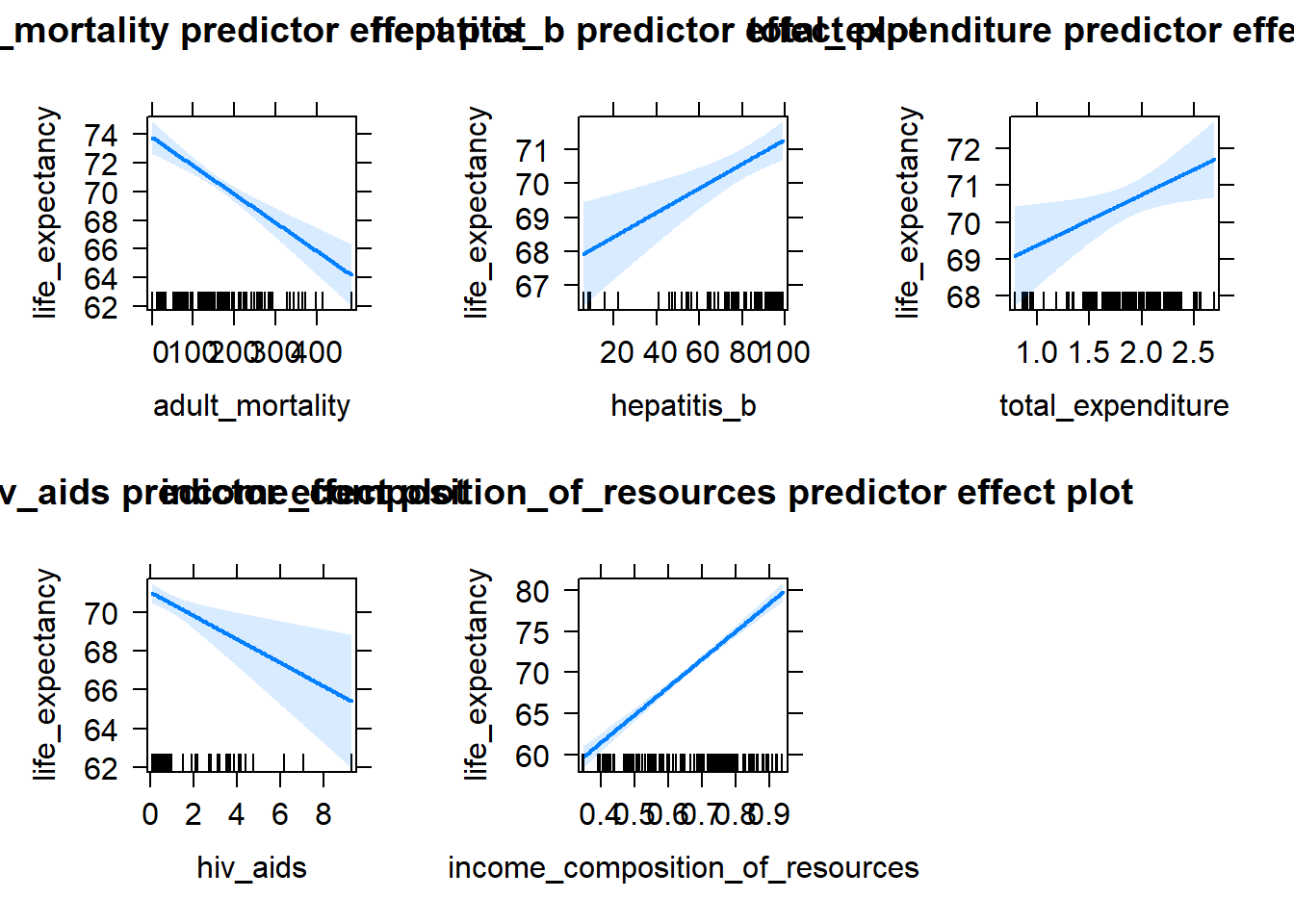
# Visų koeficientų interpretacija paprasta,  
 # nes pažingsnine regresija neišrinkti transformuoti kintamieji  
library(lm.beta)  
# Standartizuoti koeficientai  
lm.beta(model\_2)

##   
## 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.

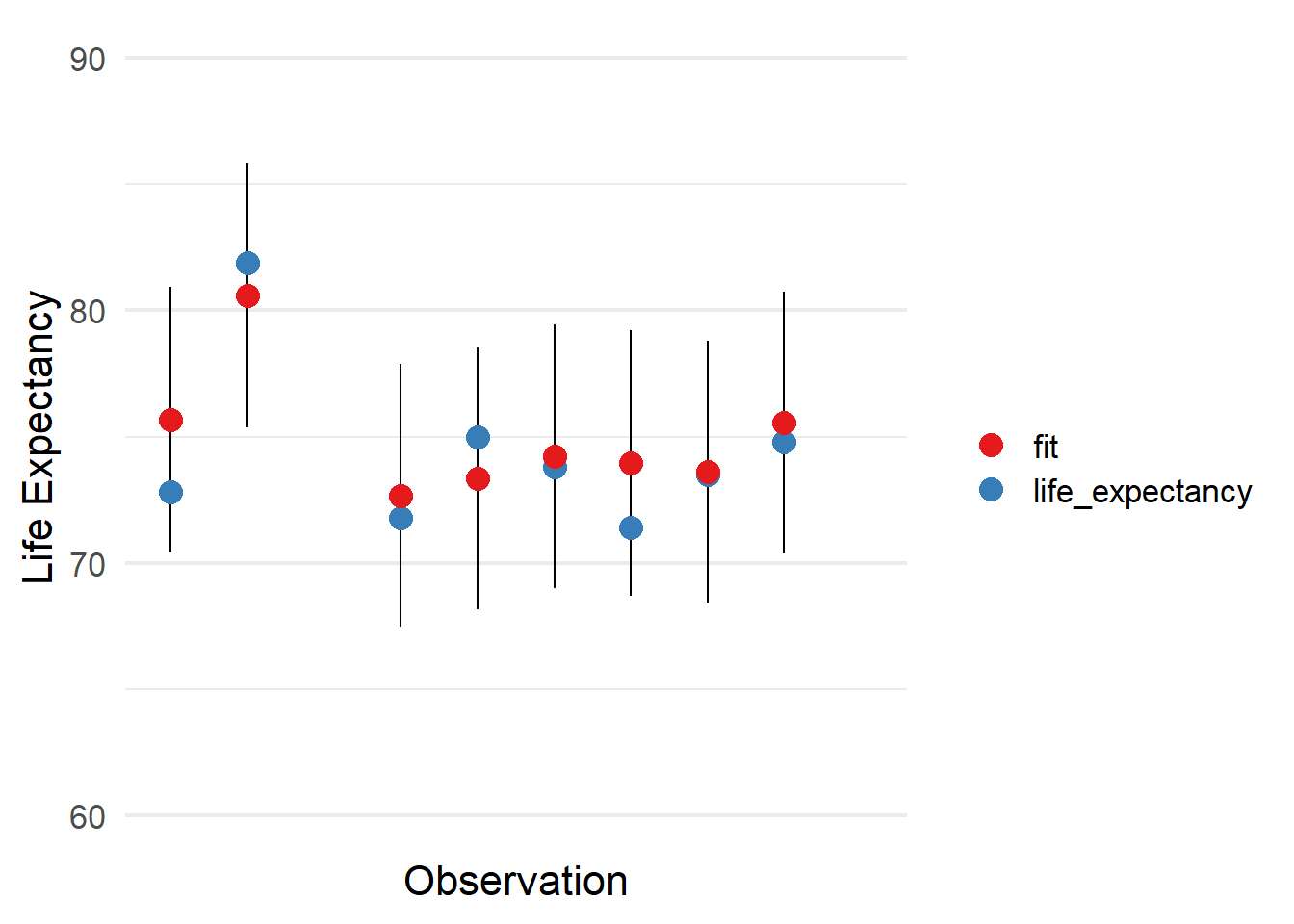
Modelio tinkamumo analizė

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-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 R2 = 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, R2 = 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));

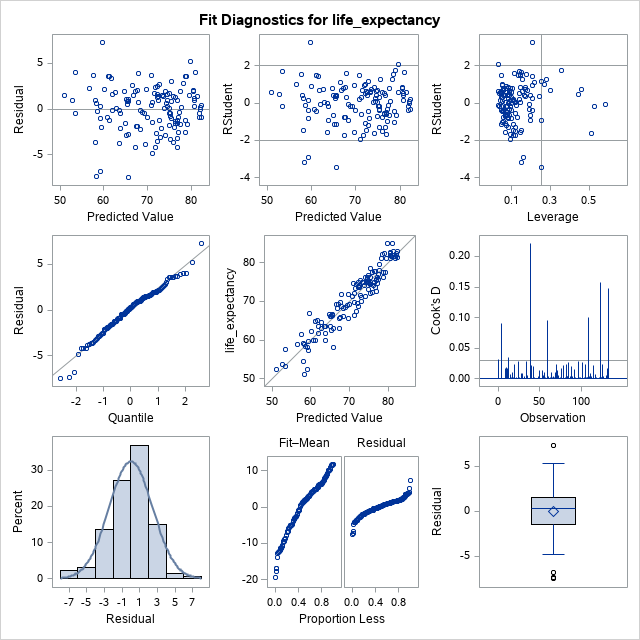
MODEL life\_expectancy = adult\_mortality infant\_deaths alcohol hepatitis\_b measles

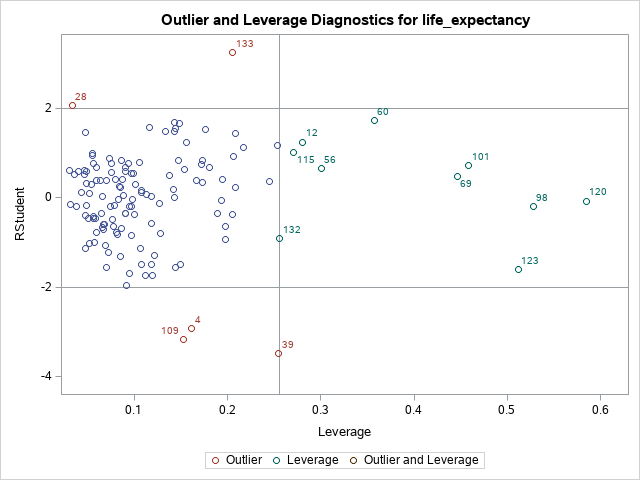
bmi under\_five\_deaths polio total\_expenditure diphtheria hiv\_aids

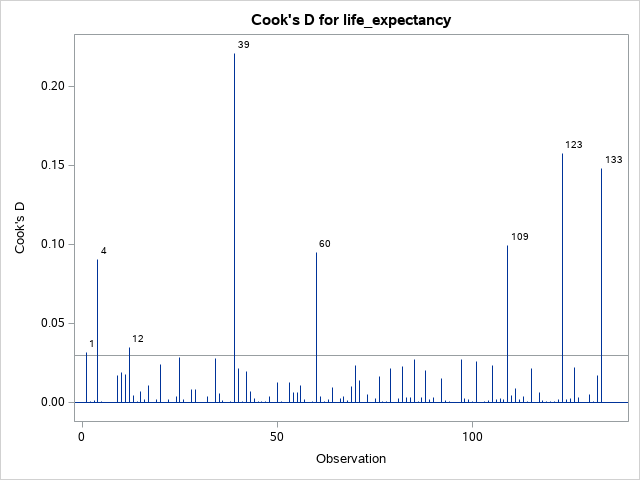
thinness\_1\_19\_years thinness\_5\_9\_years income\_composition\_of\_resources

schooling gdp;

run;







/\* 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:])

Paveikslėlis, kuriame yra langas, pastatas, elektroniniai prietaisai

Automatiškai sugeneruotas aprašymas

### 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()

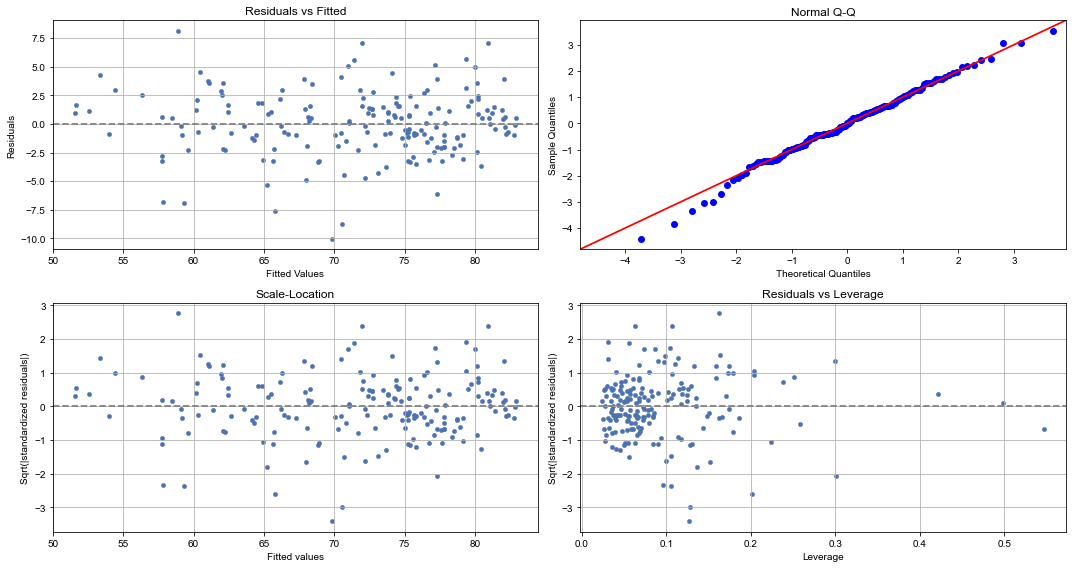
|  |
| --- |
|  |
| **Dep. Variable:** | lifeexpectancy | **R-squared:** | 0.880 |
| **Model:** | OLS | **Adj. R-squared:** | 0.869 |
| **Method:** | Least Squares | **F-statistic:** | 76.43 |
| **Date:** | Thu, 09 Dec 2021 | **Prob (F-statistic):** | 8.29e-68 |
| **Time:** | 19:55:55 | **Log-Likelihood:** | -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:])

Paveikslėlis, kuriame yra langas, pastatas

Automatiškai sugeneruotas aprašymas

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\_thinness119years',

'dfb\_thinness59years', 'dfb\_incomecompositionofresources',

'dfb\_schooling', 'dfb\_gdppercapita', 'cooks\_d', 'standard\_resid',

'hat\_diag', 'dffits\_internal', 'student\_resid', 'dffits'],

dtype='object')

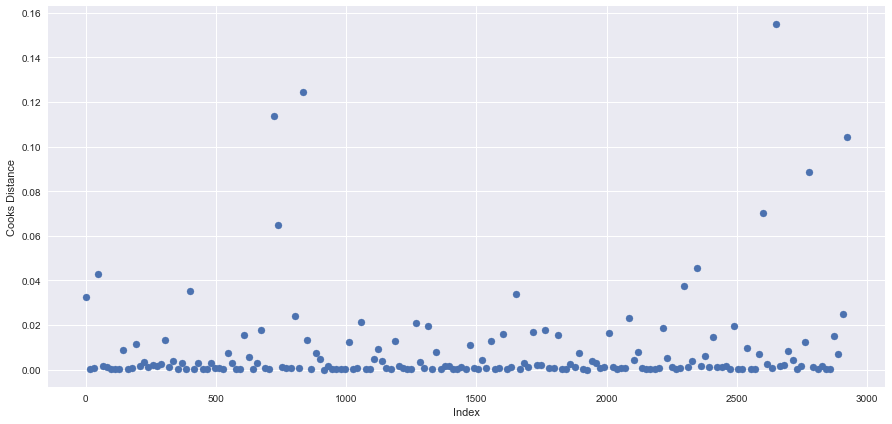
plt**.**figure(figsize**=**(15, 7))

plt**.**scatter(df**.**index, df**.**cooks\_d)

plt**.**xlabel('Index')

plt**.**ylabel('Cooks Distance')

plt**.**show()



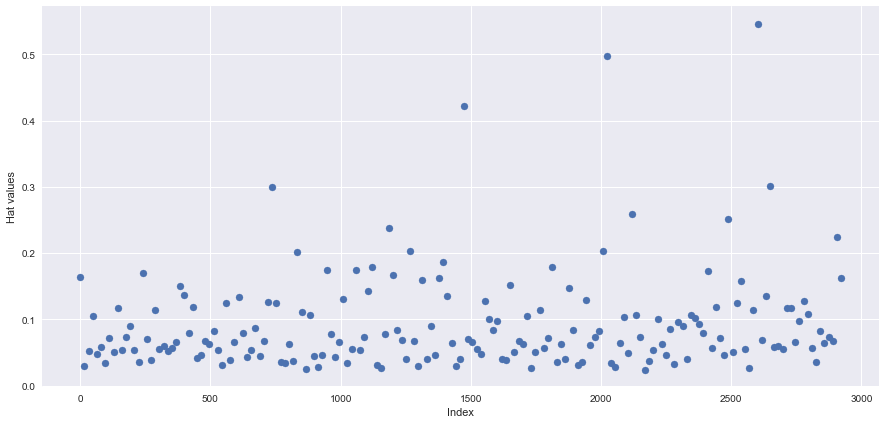
plt**.**figure(figsize**=**(15, 7))

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