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.

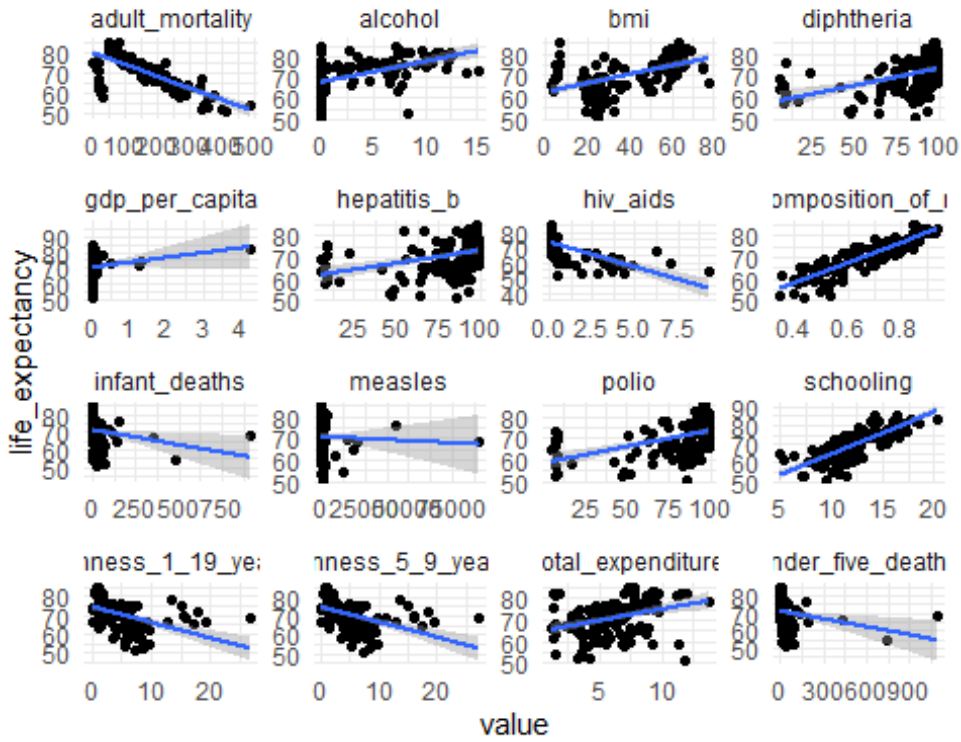
Atliktos analizės aprašymas

1. Naudojant R

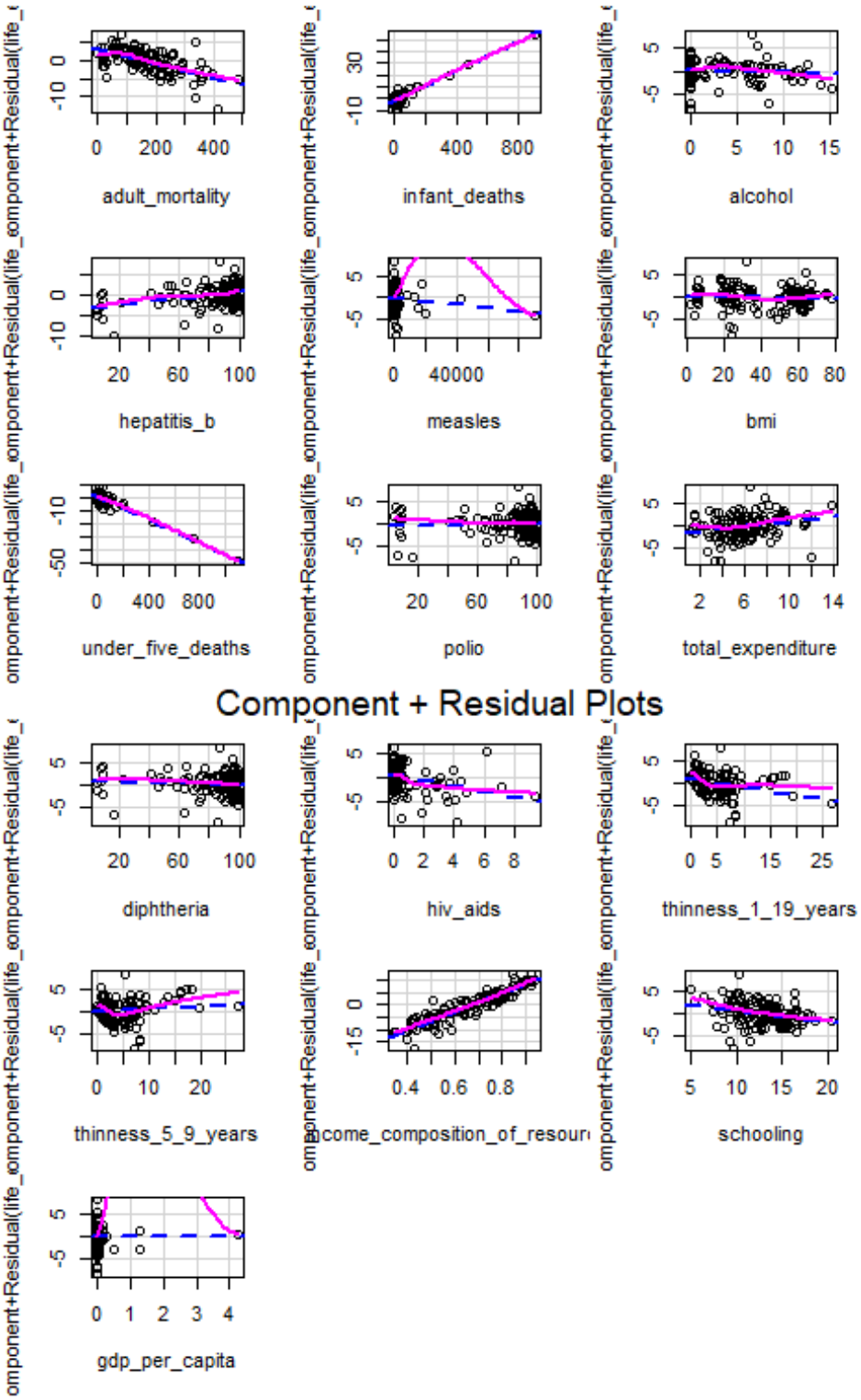
```
library(tidyverse)
library(car)
library(janitor)
x <- read_csv("life.csv") %>% clean_names()
```

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

```
set.seed(100)
transform_1<- function(x) {
  x %>%
  mutate(gdp_per_capita = gdp / population) %>%
  group_by(country) %>%
  fill(everything(), .direction = "up") %>%
  dplyr::select(-c(1, 3), -population, -gdp, -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 progrnozuoja 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()
```

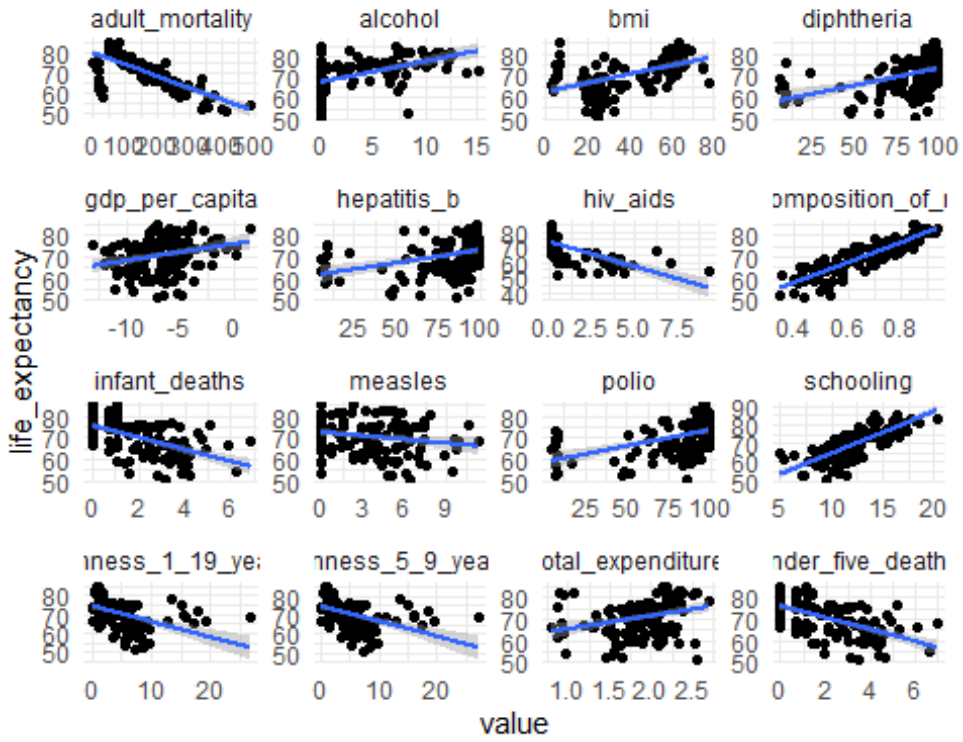



```
model <- lm(life_expectancy ~ ., data = x_1)
crPlots(model)
```




```r
transform_2 <- function(x) {
    x %>%
    mutate(gdp_per_capita = log(gdp_per_capita),
    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()
```

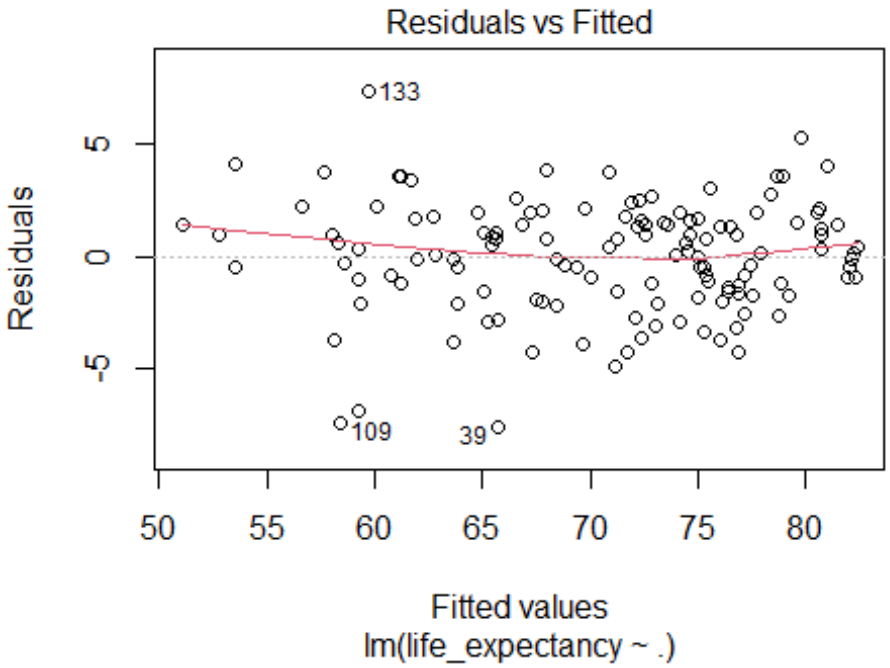


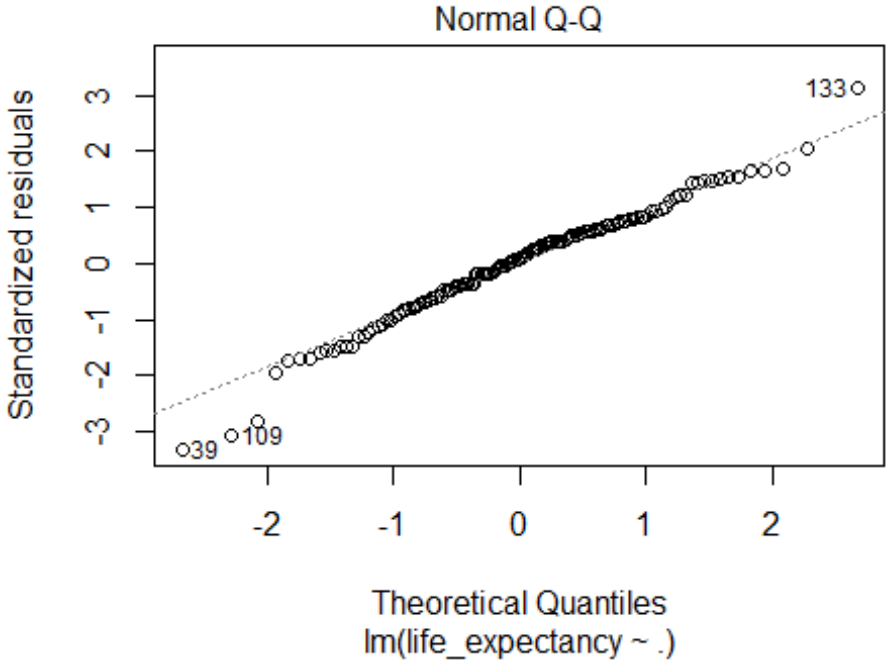
Modifikuoti duomenys išsaugomi faile „life_modified.csv“.

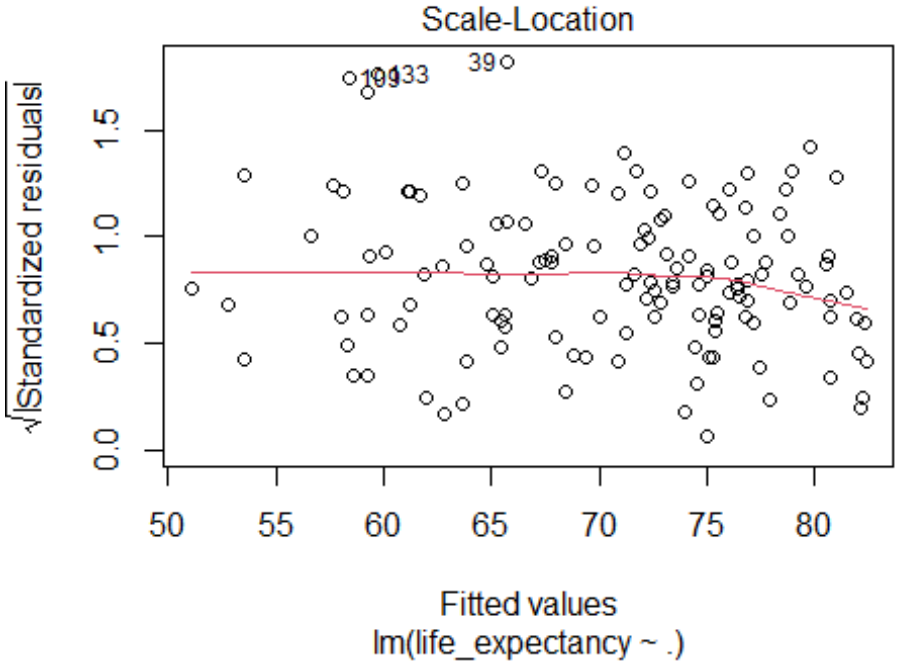
```
write.csv(x_2, "life_modified.csv")

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

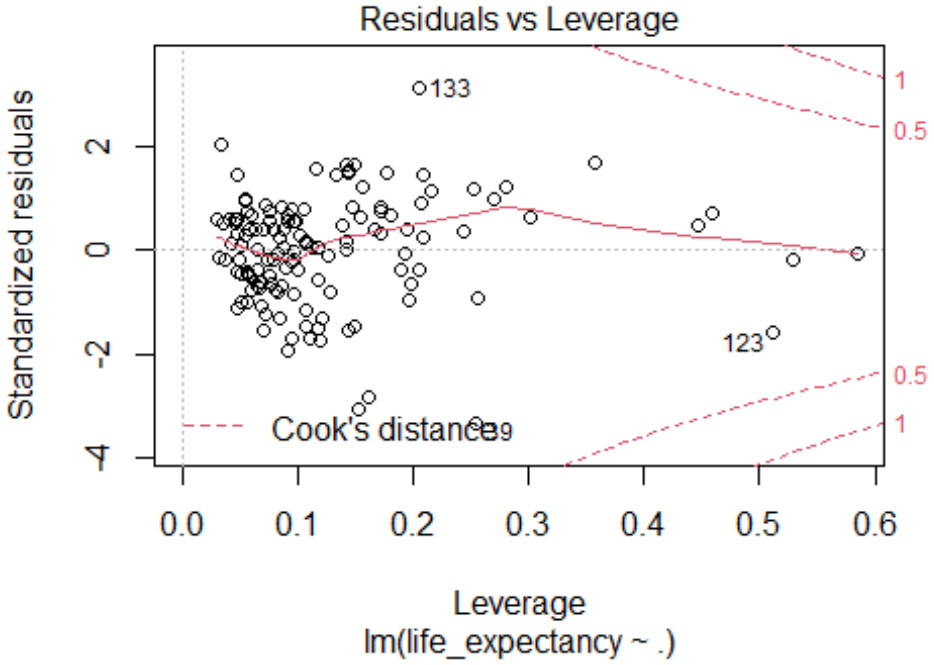
Modelio prielaidos

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

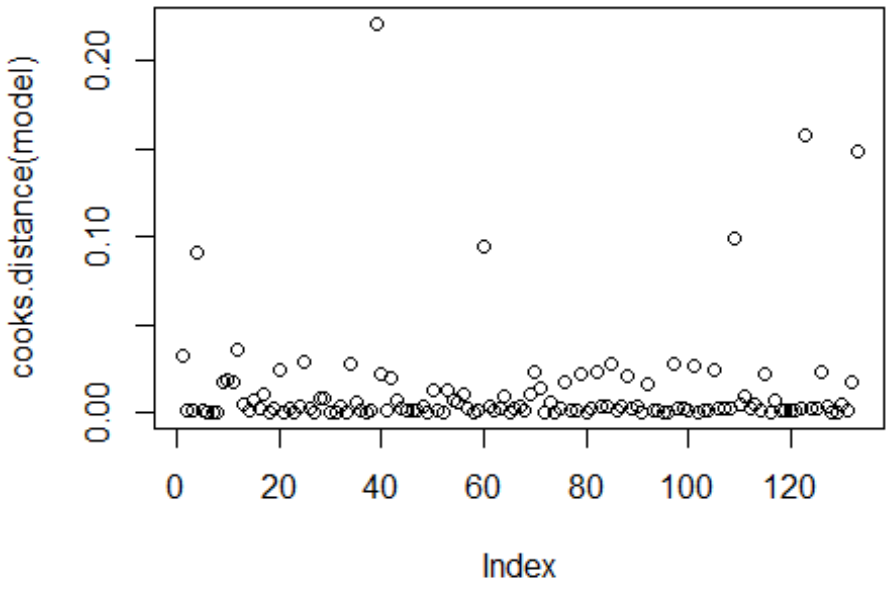

Residuals vs Fitted
Residuals
Fitted values
lm(life_expectancy ~ .)
133
109
39
50
55
60
65
70
75
80
5
0
-5


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

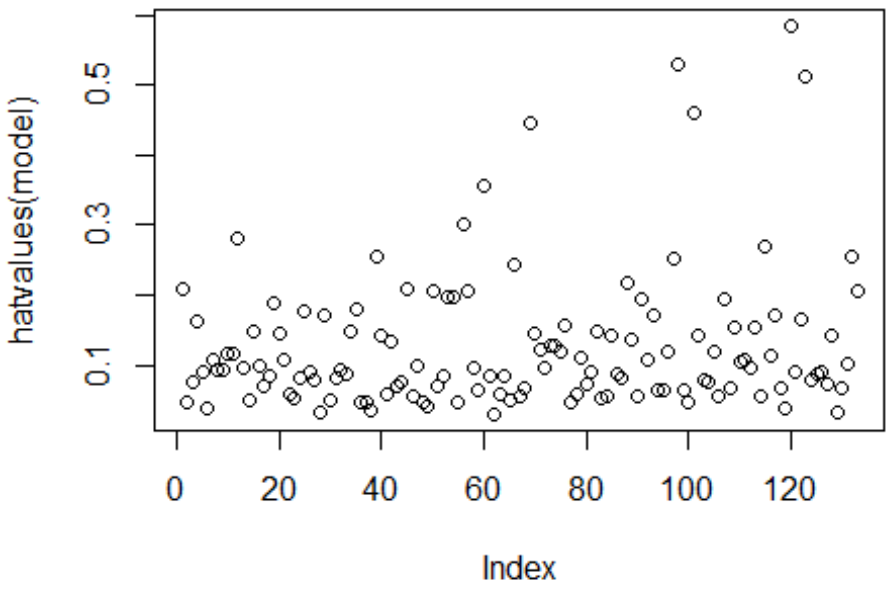





```
plot(cooks.distance(model))
```

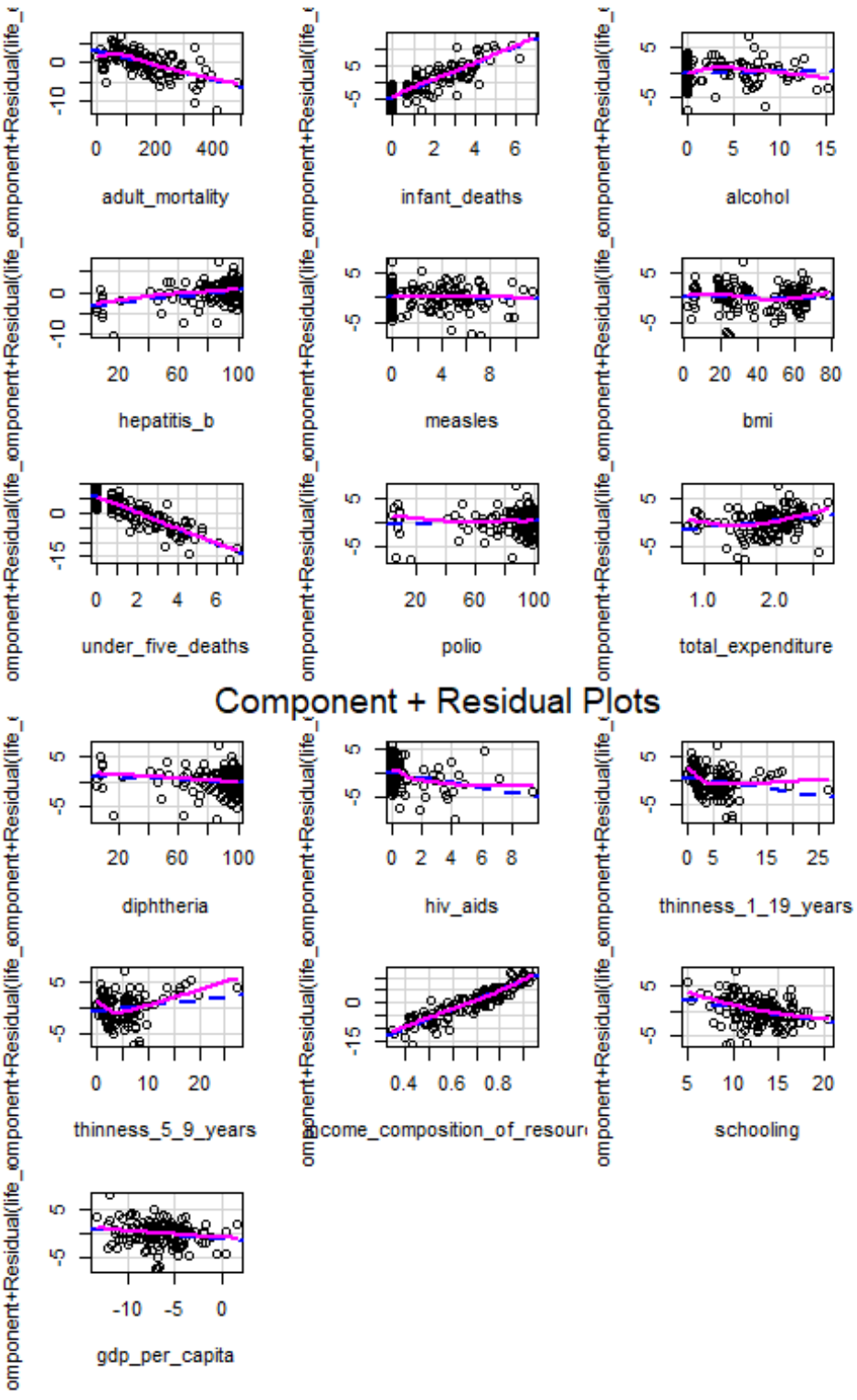



```
plot(hatvalues(model))
```




```
# 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)
```



```r
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
```

Modelio parinkimas

```r
# 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

```r
# 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
```

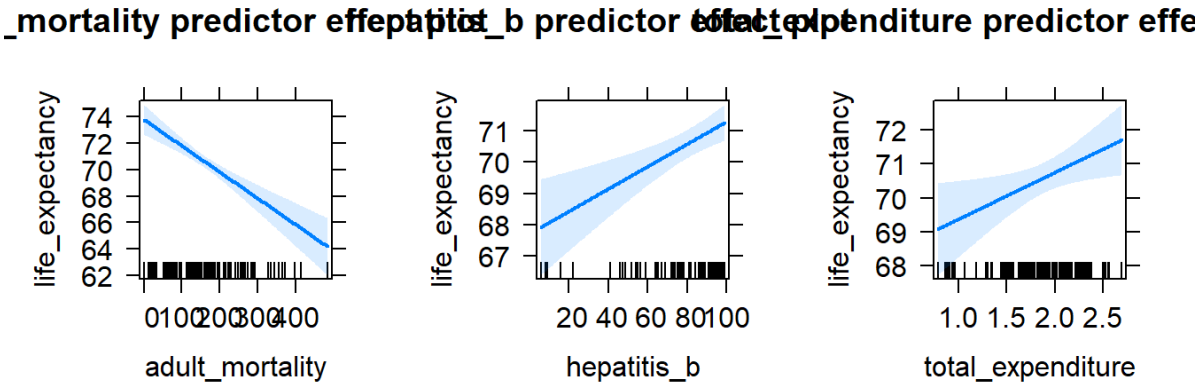
```
  # Visų koeficientų interpretacija paprasta,
  # nes pažinksnine regresija neišrinkti transformuoti kintamieji
library(lm.beta)
# Standartizuoti koeficientai
lm.beta(model_2)
```

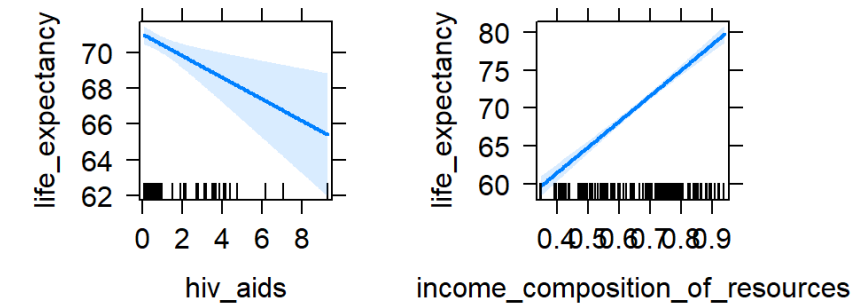
```
##
## Call:
## lm(formula = life_expectancy ~ adult_mortality + hepatitis_b +
##     total_expenditure + hiv_aids + income_composition_of_resources,
##     data = x_2)
##
## Standardized Coefficients::
##                     (Intercept)                  adult_mortality
##                      0.00000000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

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





Multikolinearumo tikrinimas

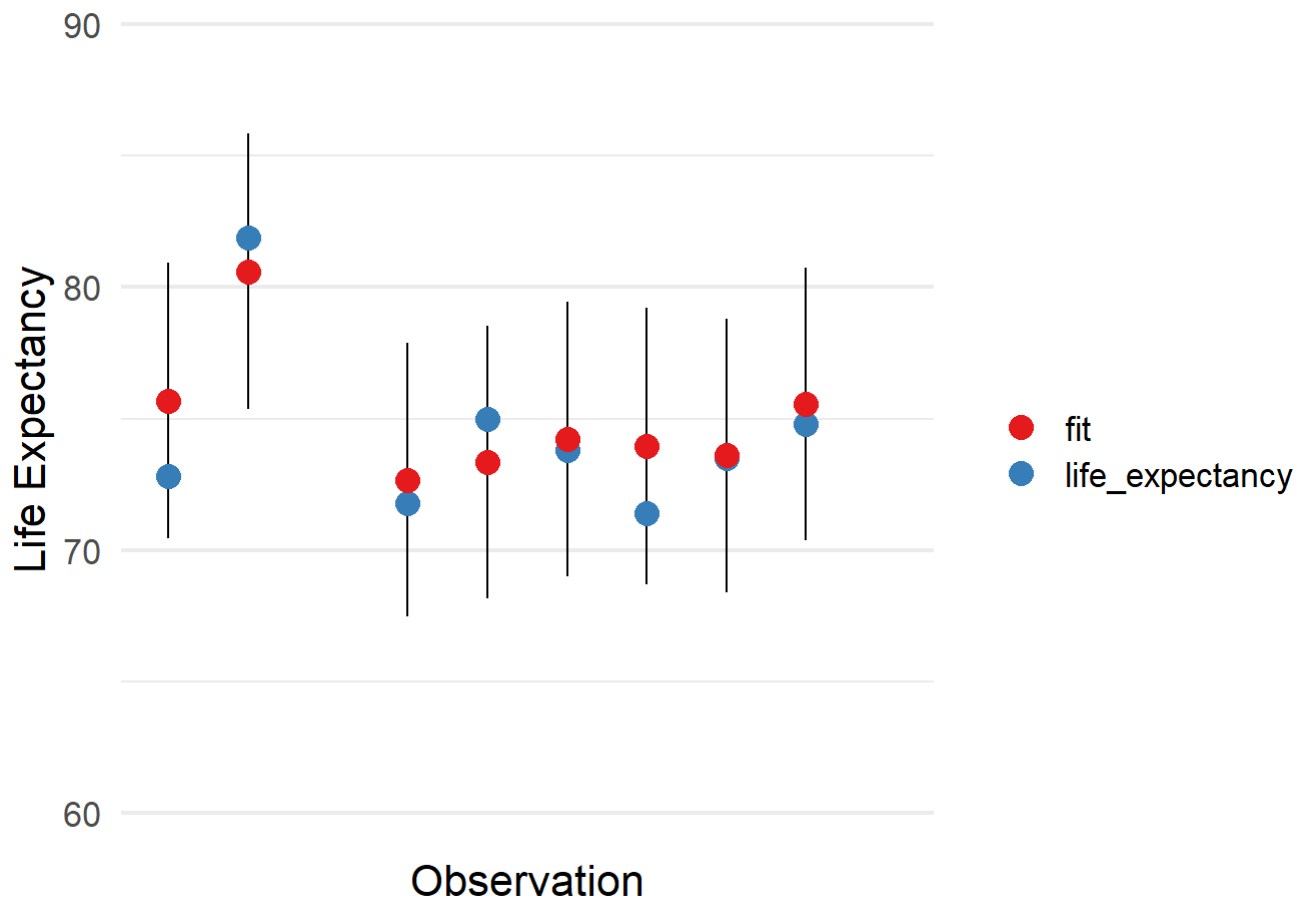
```
vars <- dplyr::select(x_2, c(adult_mortality, hepatitis_b, total_expenditure,
  hiv_aids, income_composition_of_resources, life_expectancy))

#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
```

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)
```



Rezultatai

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

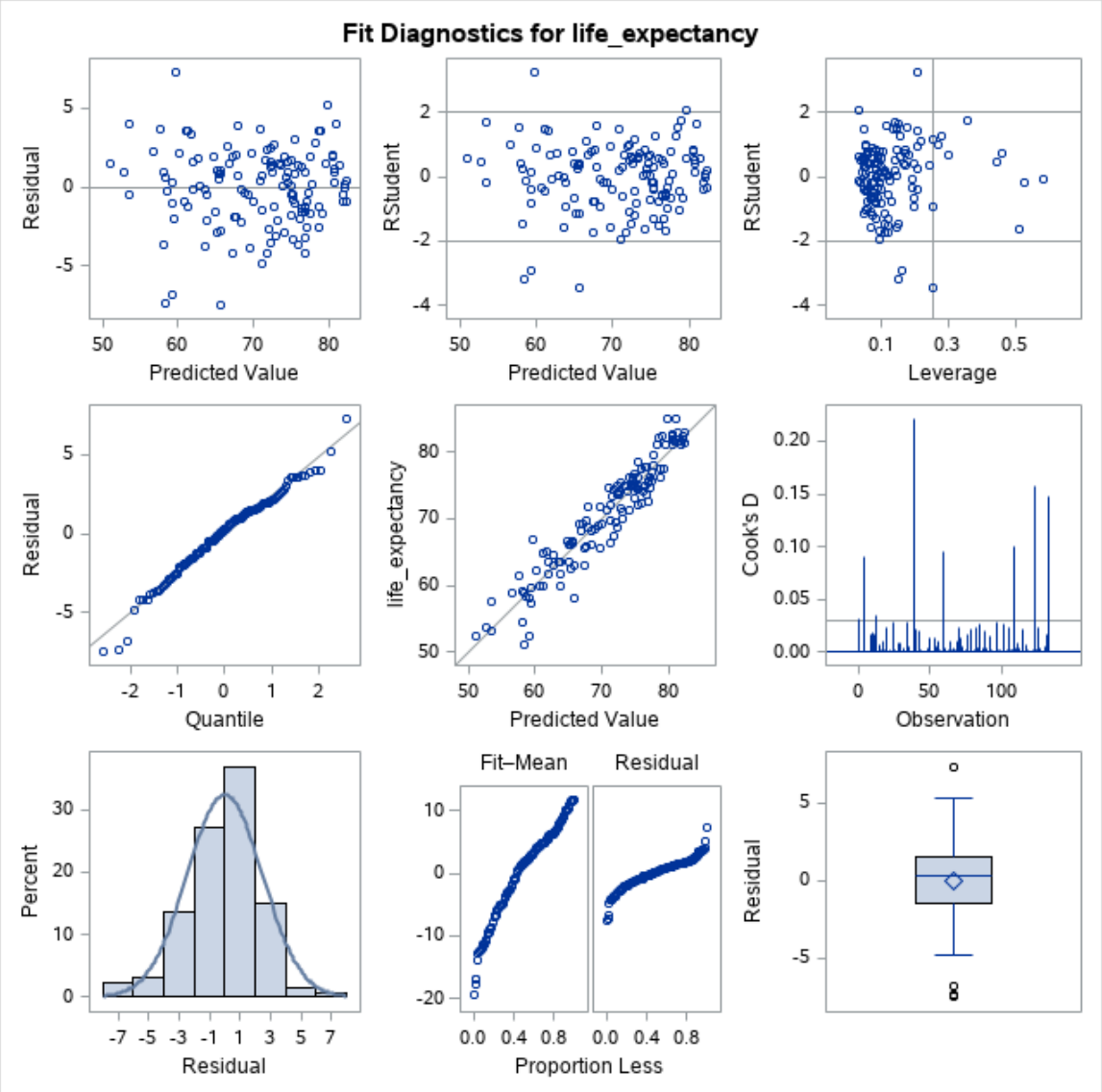
Pažinksnine regresija išrinktas modelis paaiškina 89.7% duomenų sklaidos (*F(5,127) = 221.1, R^2 = 0.897, p<0.01).* Rastos 5 statistiškai reikšmingos kovariantės gyvenimo trukmės prognozavimui (pateikti standartizuoti krypties koeficientai):

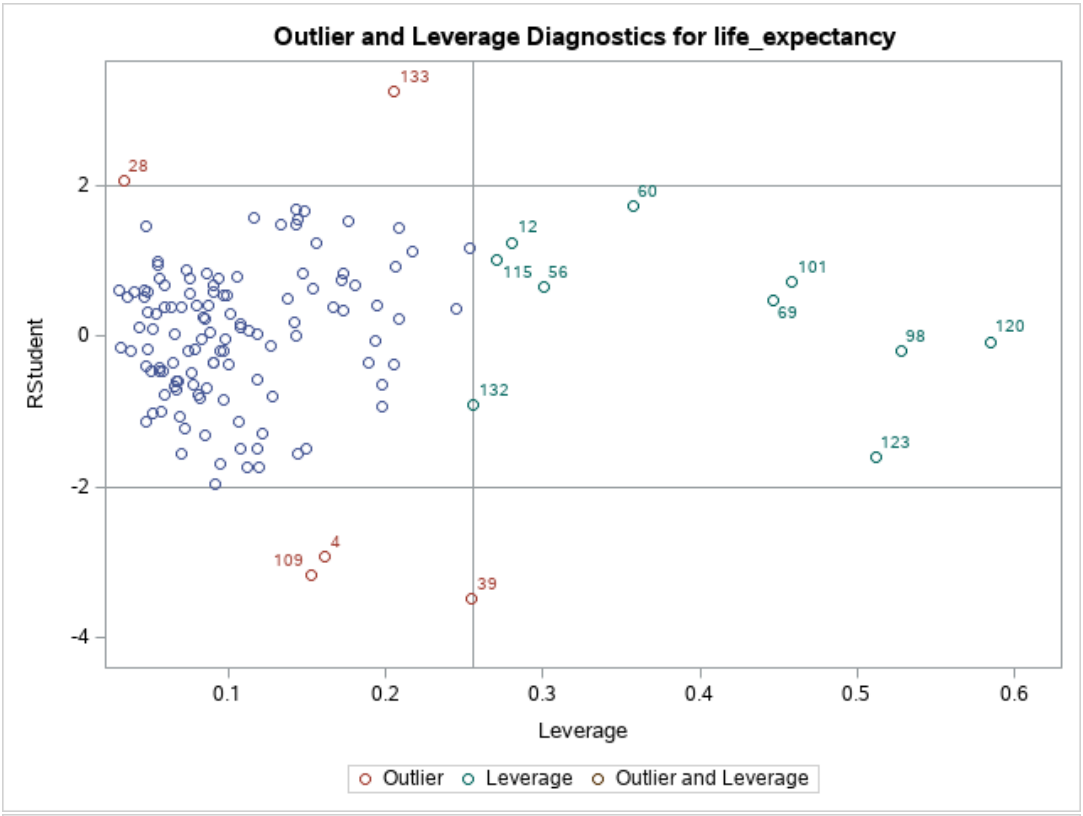
Suaugusių mirtingumas (tikimybė mirti tarp 15 ir 60 metų 1000 gyventojų) (stulp. adult_*mortality β=-0.25, p<0.001*)
Imunizacija nuo Hepatito B tarp 1 metų vaikų % (stulp. hepatitis_b *β=0.11, p<0.001*)
Dalis visų vyriausybės išlaidų sveikatos apsaugai (stulp. total_expenditure *β=0.07, p=0.02*)
Mirtys 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*)

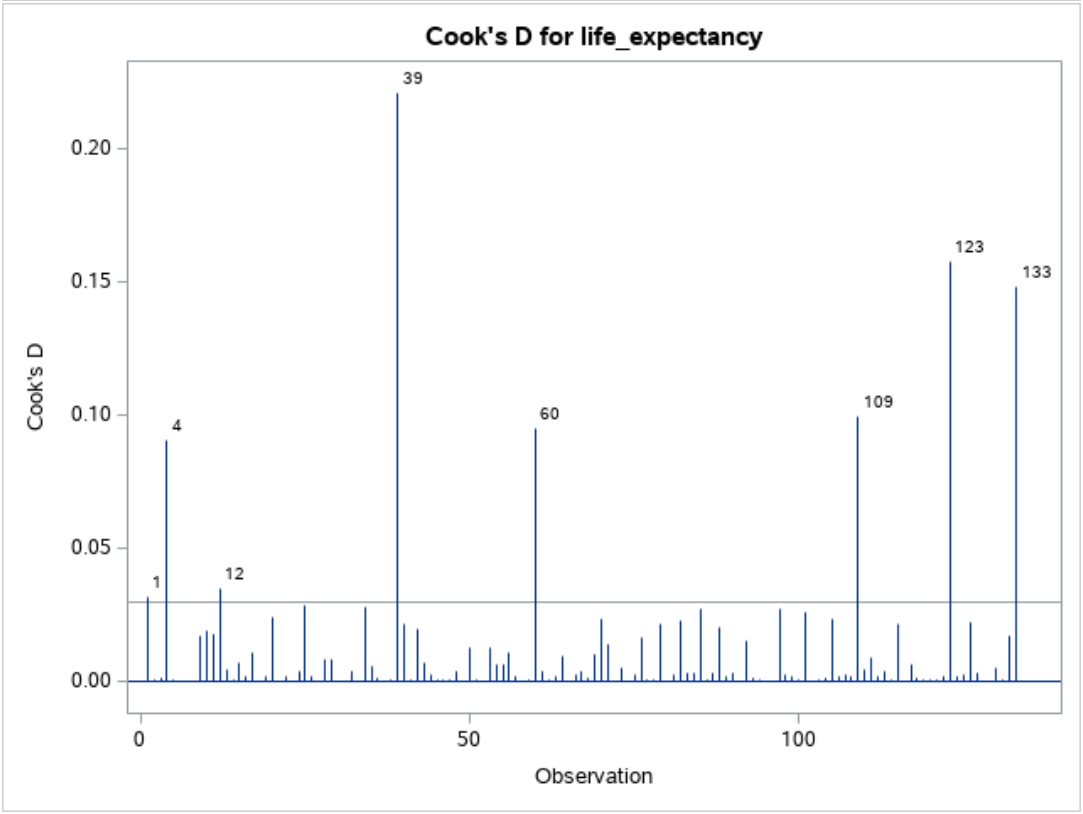
2. Naudojant SAS

Naudojamas anksčiau sukurtas duomenų failas.

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

/* 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_per_capita;
run;
```




Outlier and Leverage Diagnostics for life_expectancy
RStudent
Leverage
Outlier
Leverage
Outlier and Leverage
133
28
60
12
115
56
101
69
98
120
132
123
109
4
39


Cook's D for life_expectancy
Cook's D
Observation
39
123
133
109
60
4
12
1

```
/* 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 pažinksninę 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 gdp_per_capita / 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 total_expenditure Entered: R-Square = 0.8970 and C(p) = 4.1881

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	5	7520.59056	1504.11811	221.12	<.0001
Error	127	863.90673	6.80242		
Corrected Total	132	8384.49729			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	46.01582	1.87996	4075.49110	599.12	<.0001
adult_mortality	-0.01982	0.00328	248.41814	36.52	<.0001
hepatitis_b	0.03577	0.00989	89.00457	13.08	0.0004
total_expenditure	1.38367	0.58964	37.45889	5.51	0.0205
hiv_aids	-0.60805	0.20216	61.53858	9.05	0.0032
income_composition_of_resources	33.93718	1.94805	2064.49803	303.49	<.0001

Bounds on condition number: 2.0827, 39.209

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	0.8052	0.8052	107.730	541.34	<.0001
2	adult_mortality		2	0.0619	0.8671	34.4953	60.56	<.0001
3	hepatitis_b		3	0.0187	0.8857	13.8226	21.07	<.0001
4	hiv_aids		4	0.0068	0.8925	7.6163	8.04	0.0053
5	total_expenditure		5	0.0045	0.8970	4.1881	5.51	0.0205

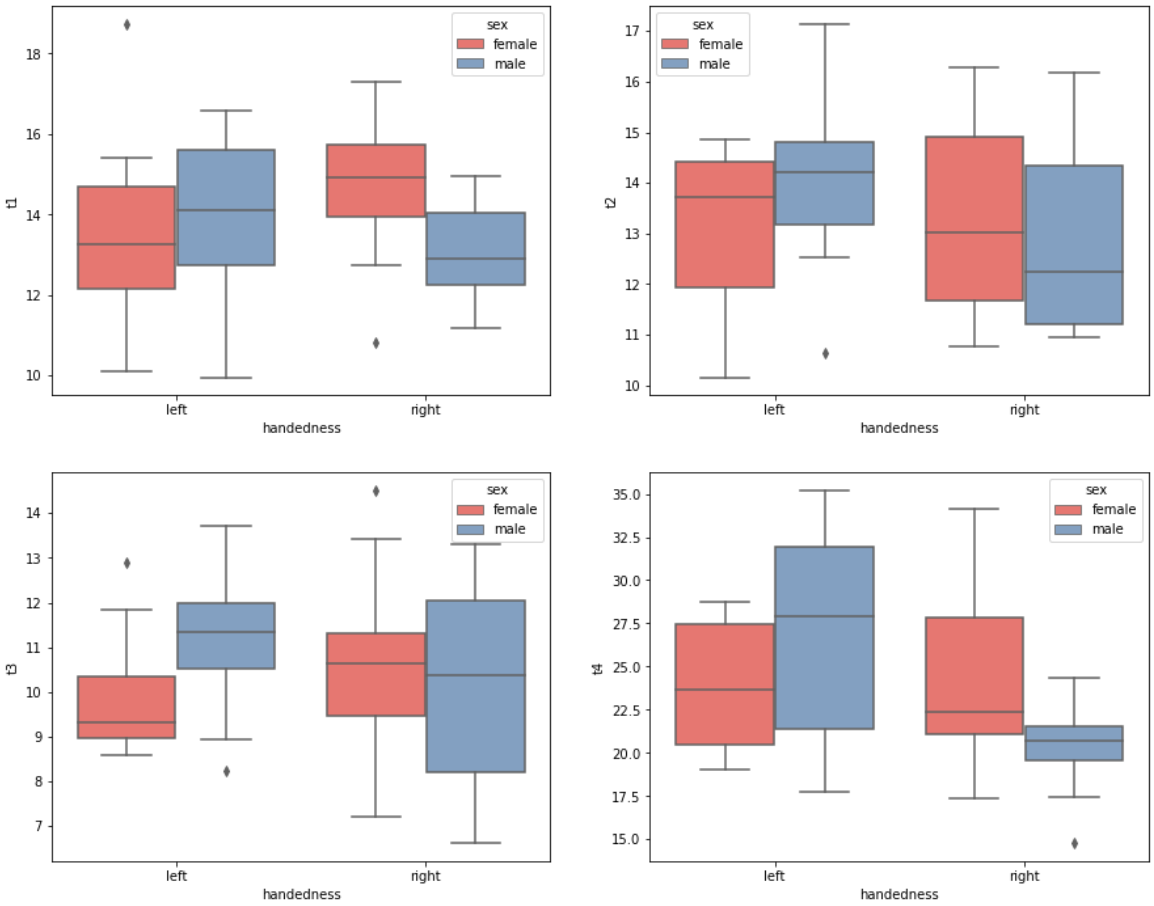
3. Naudojant Python

```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import levene
import statsmodels.api as sm
from statsmodels.formula.api import ols
import pylab
import scipy.stats as stats
from bioinfokit.analys import stat

def split(df, col):
    return [
        df[df["handedness"] == "left"][df["sex"] == "female"][col],
        df[df["handedness"] == "left"][df["sex"] == "male"][col],
        df[df["handedness"] == "right"][df["sex"] == "female"][col],
        df[df["handedness"] == "right"][df["sex"] == "male"][col]
    ]

def vartest(df,col):
    s = split(df,col)
    stat, p = levene(s[0],s[1],s[2],s[3])
    print("F value:", round(stat,4), "Pr(>F)", round(p,4))

def anova(df, col):
    stats.probplot(df[col], dist="norm", plot=pylab)
    pylab.show()
    model = ols(col + ' ~ sex * handedness', data=df).fit()
    anova_table = sm.stats.anova_lm(model, typ=3)
    return anova_table
```

```
data = pd.read_csv("data.csv")
data = data.sort_values(["sex", "handedness"])
mypal = {sex: '#f9665e' if sex == "female" else '#799fcb' for sex in data["sex"].unique()}

ft = data
ft["group"] = ft["handedness"] + ft["sex"]
```

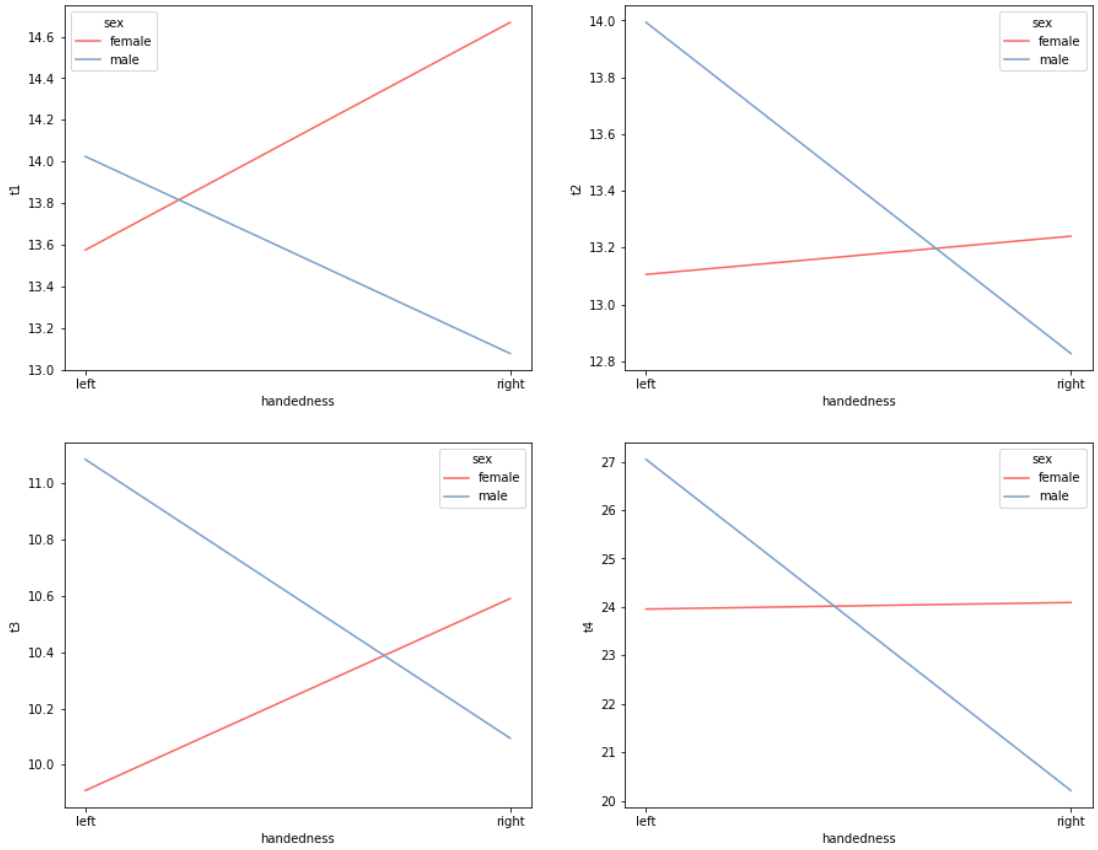
```
fig, axes = plt.subplots(2, 2,figsize=(15,12))
fig.suptitle("Tiriamieji grafikai")
sns.boxplot(ax = axes[0,0],x="handedness", y="t1", hue="sex", data=data, palette=mypal)
sns.boxplot(ax = axes[0,1],x="handedness", y="t2", hue="sex", data=data, palette=mypal)
sns.boxplot(ax = axes[1,0],x="handedness", y="t3", hue="sex", data=data, palette=mypal)
sns.boxplot(ax = axes[1,1],x="handedness", y="t4", hue="sex", data=data, palette=mypal)
```

Tiriamieji grafikai



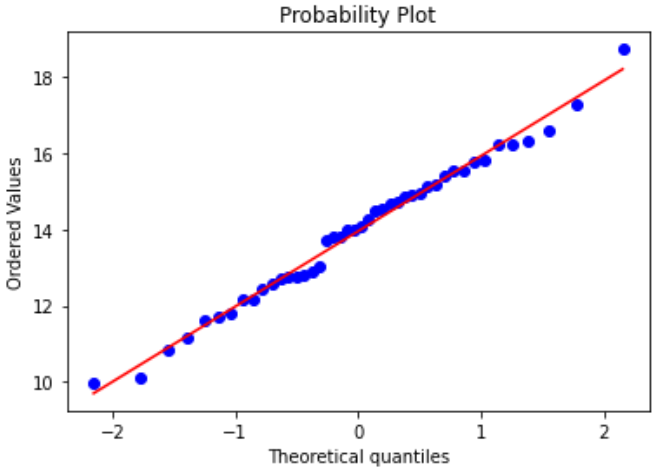

```
means = data.groupby(['sex','handedness']).mean()

fig, axes = plt.subplots(2, 2,figsize=(15,12))
fig.suptitle("Vidurkių grafikas")
sns.lineplot(ax = axes[0,0],x="handedness", y="t1", hue="sex", data=means, palette=mypal)
sns.lineplot(ax = axes[0,1],x="handedness", y="t2", hue="sex", data=means, palette=mypal)
sns.lineplot(ax = axes[1,0],x="handedness", y="t3", hue="sex", data=means, palette=mypal)
sns.lineplot(ax = axes[1,1],x="handedness", y="t4", hue="sex", data=means, palette=mypal)
```

Vidurkių grafikas

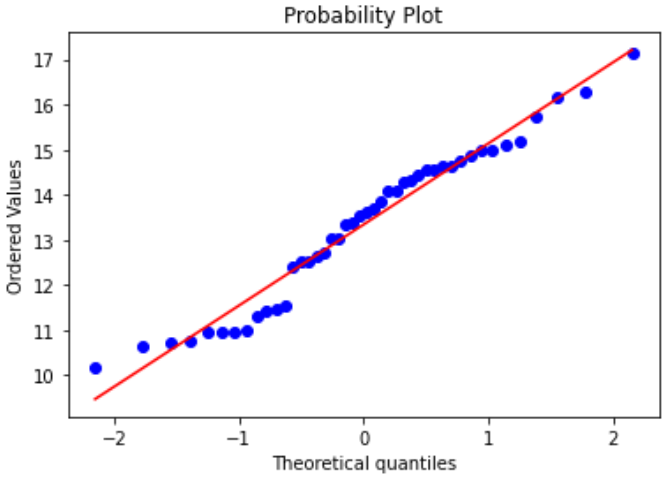



```
vartest(data,"t1")
F value: 0.8848 Pr(>F) 0.4572

vartest(data,"t2")
F value: 0.4793 Pr(>F) 0.6985

vartest(data,"t3")
F value: 1.7389 Pr(>F) 0.1745

vartest(data,"t4")
F value: 1.8124 Pr(>F) 0.1604
```

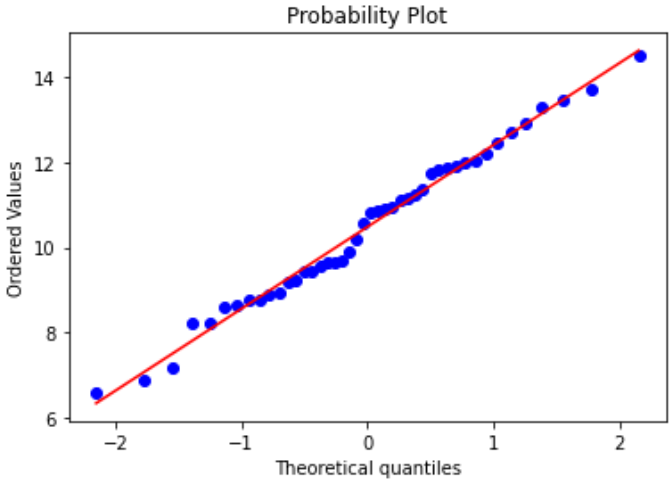
	sum_sq	df	F	PR(>F)
Intercept	1842.757078	1.0	507.629040	2.428428e-24
sex	1.102086	1.0	0.303594	5.847024e-01
handedness	6.979565	1.0	1.922679	1.732417e-01
sex:handedness	10.967304	1.0	3.021191	8.987678e-02
Residual	145.205016	40.0	NaN	NaN



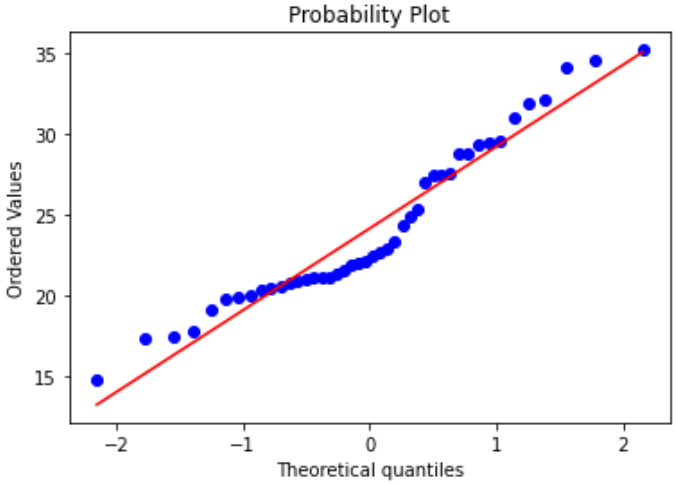
	sum_sq	df	F	PR(>F)
Intercept	1717.564555	1.0	538.700233	8.037321e-25
sex	4.301995	1.0	1.349286	2.522877e-01
handedness	0.105805	1.0	0.033185	8.563713e-01
sex:handedness	4.459198	1.0	1.398591	2.439417e-01
Residual	127.533975	40.0	NaN	NaN



	sum_sq	df	F	PR(>F)
Intercept	981.812039	1.0	276.846236	1.413343e-19
sex	7.541214	1.0	2.126432	1.525851e-01
handedness	2.710706	1.0	0.764351	3.871895e-01
sex:handedness	7.358940	1.0	2.075035	1.575086e-01
Residual	141.856657	40.0	NaN	NaN



	sum_sq	df	F	PR(>F)
Intercept	5738.425909	1.0	261.107038	3.929689e-19
sex	52.051729	1.0	2.368432	1.316855e-01
handedness	0.108554	1.0	0.004939	9.443204e-01
sex:handedness	127.945606	1.0	5.821718	2.050468e-02
Residual	879.091725	40.0	NaN	NaN




```
res = stat()
res.tukey_hsd(df=ft, res_var='t4', xfac_var='group', anova_model='t4 ~ group')
res.tukey_summary
```

	group1	group2	Diff	Lower	Upper	q-value	p-value
0	leftfemale	rightfemale	0.136416	-5.066654	5.339486	0.099392	0.900000
1	leftfemale	leftmale	3.089145	-2.291556	8.469846	2.176434	0.425942
2	leftfemale	rightmale	3.745010	-2.215856	9.705877	2.381714	0.345613
3	rightfemale	leftmale	2.952729	-1.990949	7.896407	2.264224	0.390736
4	rightfemale	rightmale	3.881426	-1.688128	9.450981	2.641903	0.257923
5	leftmale	rightmale	6.834156	1.098309	12.570002	4.516826	0.013984