Tugas Kelompok Analisis Regresi Kelompok 3

Muhammad Haikal Rasyadan, Raihana Asma Amani, Delita Nur Hasanah

2024-02-08

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggridges)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(plotly)
```

```
##
## Attaching package: 'plotly'
##
## The following object is masked from 'package:ggplot2':
##
##     last_plot
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following object is masked from 'package:graphics':
##
##     layout
```

Data

Analisis ini bertujuan untuk mengetahui hubungan antara umur (x) dan kadar kolesterol (y) dalam memperkirakan risiko penyakit jantung. Umur dipilih sebagai variabel independen karena diasumsikan berkorelasi

dengan kadar kolesterol seiring bertambahnya usia, sementara kadar kolesterol dipilih karena perannya yang signifikan dalam mengevaluasi risiko penyakit jantung.

Analisis menggunakan model regresi linier sederhana untuk memahami dampak umur terhadap kadar kolesterol, dengan tujuan memperoleh wawasan lebih lanjut tentang hubungan tersebut dan implikasinya terhadap kesehatan.

```
data <- read.csv("/Users/user/Downloads/Documents/Anreg /Heart_Disease_Prediction.csv", sep=",")
y<-data$Cholesterol
x<-data$Age
n <- nrow(data)
n
```

```
## [1] 270
```

```
data<-data.frame(cbind(y,x))
head(data)
```

```
##     y  x
## 1 322 70
## 2 564 67
## 3 261 57
## 4 263 64
## 5 269 74
## 6 177 65
```

Pembentukan Model

```
model <- lm(y~x, data)
model
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Coefficients:
## (Intercept)            x
##     181.692        1.249
```

```
summary(model)
```

```
##
## Call:
## lm(formula = y ~ x, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -126.864  -34.233   -3.756   28.697  298.650
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 181.6920    18.6593   9.737  < 2e-16 ***
## x             1.2486     0.3381   3.693 0.000269 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50.51 on 268 degrees of freedom
## Multiple R-squared:  0.04842,    Adjusted R-squared:  0.04487
## F-statistic: 13.64 on 1 and 268 DF,  p-value: 0.0002686
```

Berdasarkan pemodelan dengan fungsi lm, didapatkan estimasi persamaan regresi linier yang juga dapat disebut sebagai nilai perkiraan dari variabel respons Y, sebagai berikut:

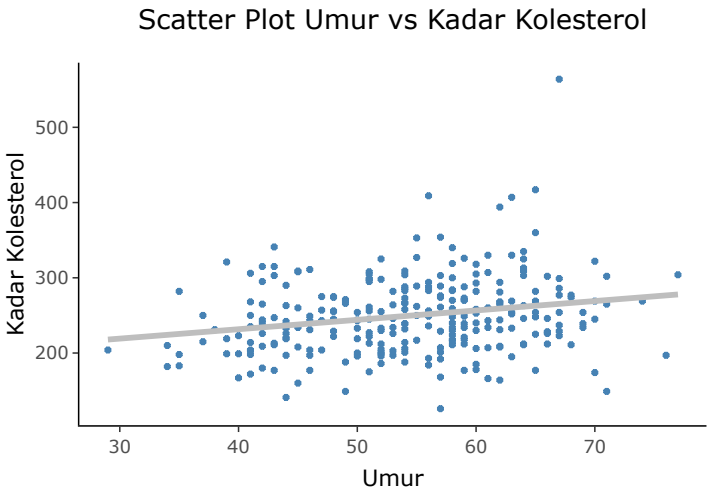
$$E[\hat{Y}] = \hat{Y} = 181.6920 + 1.2486x$$

Hasil regresi menunjukkan adanya hubungan positif antara umur dan kadar kolesterol, yang mengindikasikan bahwa semakin panjang umur seseorang, semakin tinggi kemungkinannya memiliki kadar kolesterol yang lebih tinggi juga. Intersep pada nilai 181,6920 menunjukkan bahwa jika umur adalah 0 (baru lahir), maka kadar kolesterol seseorang diprediksi memiliki nilai sebesar 181,6920. Kemiringan sebesar 1,2486 diduga mengindikasikan bahwa setiap bertambahnya umur akan diikuti dengan peningkatan sebesar 1,2486 rata-rata kadar kolesterol seseorang.

Visualisasi Scatter Plot

```
y.bar <- mean(y)
interactive.plot <- ggplot(model) +
  geom_point(aes(x = x,y = y),color="steelblue",shape=8, size=1) +
  geom_smooth(aes(x = x, y = y), method = "lm", se = FALSE, color = "grey") +
  ggtitle("Scatter Plot Umur vs Kadar Kolesterol") +
  ylab("Kadar Kolesterol") +
  xlab("Umur") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))
ggplotly(interactive.plot)
```

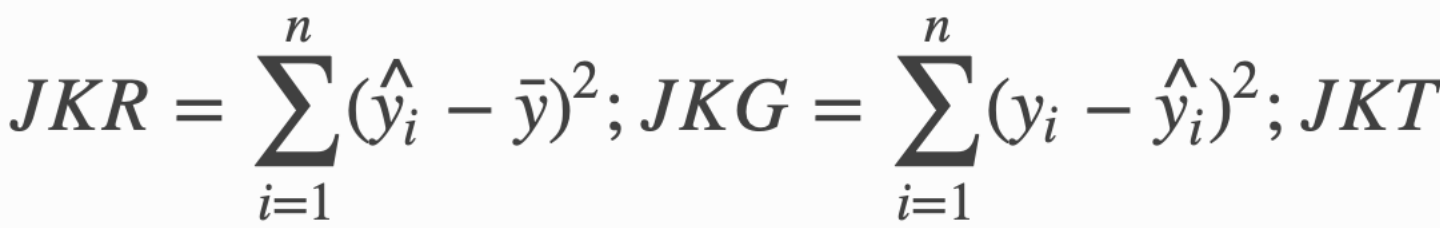
```
## `geom_smooth()` using formula = 'y ~ x'
## PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, pleas
```


Scatter Plot Umur vs Kadar Kolesterol
Kadar Kolesterol
500
400
300
200
30
40
50
60
70
Umur

Penguraian Keragaman

Dari scatter plot diatas, terlihat adanya penyimpangan relatif dari nilai harapan untuk setiap pengamatan. penyimpangan relatif ini dikenal sebagai galat. keragaman dari galat untuk setiap pengamatan dapat diuraikan berdasarkan garis dugaan persamaan ($\hat{Y}$) dan garis rata-rata respons ($\bar{Y}$). Penguraian keragaman ini terbagi menjadi Jumlah Kuadrat Regresi (JKR), Jumlah Kuadrat Galat (JKG), dan Jumlah Kuadrat Total (JKT), dengan perhitungan yang sesuai.

```
knitr::include_graphics("/Users/user/Downloads/penguraian keragaman.png")
```



$$JKR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2; JKG = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2; JKT$$

Hubungan antara ketiganya dapat pula dituliskan sebagai berikut.

$$JKT = JKR + JKG$$

Nilai penguraian keragaman ini dapat dilihat menggunakan fungsi anova, sebagai berikut:

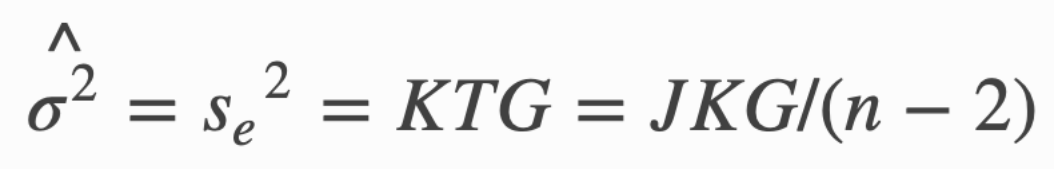
```
(anova.model <- anova(model))
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value    Pr(>F)
## x           1  34799   34799  13.638 0.0002686 ***
## Residuals 268 683825    2552
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

berdasarkan perhitungan ANOVA diatas, dapat diketahui nilai Jumlah Kuadrat Rataan (JKR) sebesar 34799 dan Jumlah Kuadrat Galat (JKG) sebesar 683825.

Ragam Galat/Error

```
knitr::include_graphics("/Users/user/Downloads/ragam galat.png")
```



$$\hat{\sigma}^2 = s_e{}^2 = KTG = JKG/(n-2)$$

```
(KTG <- anova.model$`Mean Sq`[2])
```

```
## [1] 2551.587
```

Galat Baku

Oleh karena simpangan baku merupakan akar kuadrat dari ragam, maka nilai dugaan galat baku model yang kita bentuk adalah:

```
(galat.baku <- sqrt(KTG))
```

```
## [1] 50.51324
```

Berdasarkan perhitungan diatas didapatkan hasil sebagai berikut:

$JKR = 34799$

$JKG = 683825$

$KTG = 2551.587$

$s_e = 50.51324$

Keragaman Dugaan Parameter

Dugaan Parameter β_1

Hipotesis uji:

H_0:$\beta_1 = 0$ (Kadar kolesterol tidak memiliki hubungan hubungan linear dengan faktor umur)

H_1:$\beta_1 \neq 0$ (Kadar kolesterol memiliki hubungan hubungan linear dengan faktor umur)

```
b1<-(sum(x*y)-sum(x)*sum(y)/n)/(sum(x^2)-(sum(x)^2/n))
b1
```

```
## [1] 1.248633
```

Dengan nilai slope (b1) sebesar 1.248633, selanjutnya akan dilakukan uji apakah faktor umur (x) memiliki pengaruh yang signifikan terhadap kadar kolesterol (y) dalam hubungan linier atau tidak.

```
knitr::include_graphics("/Users/user/Downloads/b1.png")
```

$$s_{\hat{\beta}_1} = \sqrt{\frac{KTG}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} ; t_{hitung} = \frac{\hat{\beta}_1 - }{s_{\hat{\beta}_1}}$$

Nilai $S\beta_1$ dapat dihitung dengan sintaks sebagai berikut.

```r
(se_b1 <- sqrt(KTG/sum((x-mean(x))^2)))
```

```
## [1] 0.3381078
```

```r
(t_b1 <- b1/se_b1)
```

```
## [1] 3.693003
```

statistik uji:

```r
p_valueb1 <- 2*pt(-abs(t_b1 ),df<-n-2)
p_valueb1
```

```
## [1] 0.0002685531
```

Kaidah Keputusan:

Karena p-value(0.0002685531) < 0.05 maka tolak H0. Oleh karena itu, hal ini menunjukkan terdapat cukup bukti untuk menyatakan adanya hubungan linier antara faktor umur (x) dan kadar kolesterol (y). Hal tersebut juga memberikan bukti yang cukup untuk menyatakan bahwa faktor umur memengaruhi kadar kolesterol pada taraf nyata 5%.

Dugaan Parameter β_0

Hipotesis uji:

H_0:$\beta_0 = 0$ (Semua nilai kadar kolesterol dapat dijelaskan oleh faktor umur)

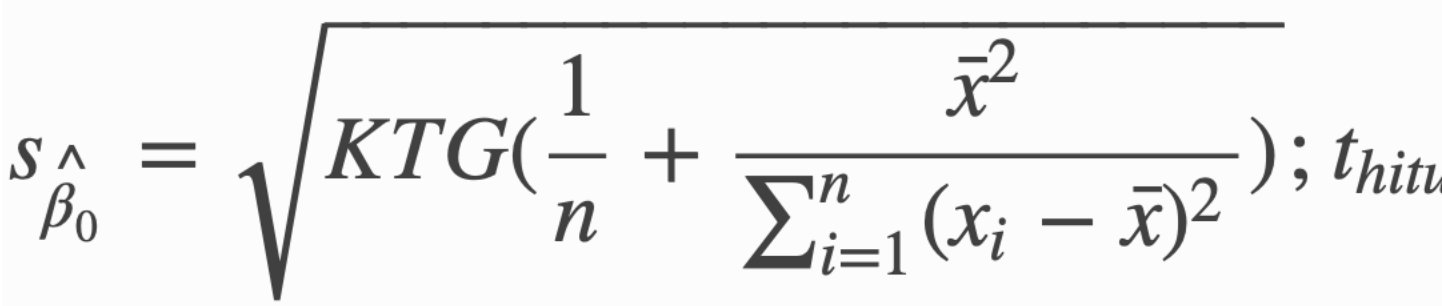
H_1:$\beta_0 \neq 0$ (Ada nilai kadar kolesterol yang tidak dapat dijelaskan oleh faktor umur)

```r
b0<-mean(y)-b1*mean(x)
b0
```

```
## [1] 181.692
```

Dengan menggunakan data yang sama bersama dengan nilai intersep (b0) sebesar 181.692, akan dilakukan uji apakah ada variasi dalam Kadar Kolesterol (y) yang tidak dapat dijelaskan oleh faktor Umur (x).

```r
knitr::include_graphics("/Users/user/Downloads/b0.png")
```



$$s_{\hat{\beta}_0} = \sqrt{KTG\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}; t_{hitu}$$

```r
(se_b0 <- sqrt(KTG*(1/n+mean(x)^2/sum((x-mean(x))^2))))
```

```
## [1] 18.65931
```

```r
(t_b0 <- b0/se_b0)
```

```
## [1] 9.737337
```

statistik uji:

```r
p_valueb0 <- 2*pt(-abs(t_b0 ),df<-n-2)
p_valueb0
```

```
## [1] 2.221061e-19
```

Kaidah Keputusan:

Karena p-value(2.221061e-19) < 0.05 maka tolak H0. Oleh karena itu, hal ini menunjukkan terdapat cukup bukti untuk menyatakan adanya nilai kadar kolesterol (y) yang tidak dapat dijelaskan oleh faktor umur (x) pada taraf nyata 5%.

Uji Korelasi dan Koefisien Determinasi

Uji Korelasi:

```r
r<-(sum(x*y)-sum(x)*sum(y)/n)/
sqrt((sum(x^2)-(sum(x)^2/n))*(sum(y^2)-(sum(y)^2/n)))
r
```

```
## [1] 0.2200563
```

Uji Koefisien Determinasi:

```r
Koef_det<-r^2
Koef_det
```

```
## [1] 0.04842478
```

Dengan korelasi sebesar 0.2200563 dan koefisien determinasi sebesar 0.04842478, terdapat hubungan positif yang lemah antara faktor umur dan kadar kolesterol serta sekitar 4.842478% variasi dalam kadar kolesterol dapat dijelaskan oleh variasi dalam faktor umur.

Penduga Selang Kepercayaan Parameter Model

```r
knitr::include_graphics("/Users/user/Downloads/sk.png")
```

$$\hat{\beta}_0 - t_{(n-2;\frac{\alpha}{2})} s_{\hat{\beta}_0} < \hat{\beta}_0 < \hat{\beta}_0 + t_{(n-2;\frac{\alpha}{2})} s_{\hat{\beta}_0}$$

$$\hat{\beta}_1 - t_{(n-2;\frac{\alpha}{2})} s_{\hat{\beta}_1} < \hat{\beta}_1 < \hat{\beta}_1 + t_{(n-2;\frac{\alpha}{2})} s_{\hat{\beta}_1}$$

Penduga selang kepercayaan 95% bagi β_1:

```
#Batas Atas beta_1
(ba.b1 <- b1 + abs(qt(0.025, df=n-2))*se_b1)
```

```
## [1] 1.914318
```

```
#Batas Bawah beta1
(bb.b1 <- b1 - abs(qt(0.025, df=n-2))*se_b1)
```

```
## [1] 0.5829478
```

Penduga selang kepercayaan 95% bagi β_0:

```
#Batas Atas beta_0
(ba.b0 <- b0 + abs(qt(0.025, df=n-2))*se_b0)
```

```
## [1] 218.4295
```

```
#Batas Bawah beta0
(bb.b0 <- b0 - abs(qt(0.025, df=n-2))*se_b0)
```

```
## [1] 144.9545
```

Sehingga dapat disusun suatu selang kepercayaan untuk β_0 dan β_1 sebagai berikut:

0.5829478< β_1 <1.914318

Yang dapat dimaknai bahwa dalam taraf kepercayaan 95%, diyakini bahwa dugaan parameter β_1 berada dalam selang 0.5829478 hingga 1.914318

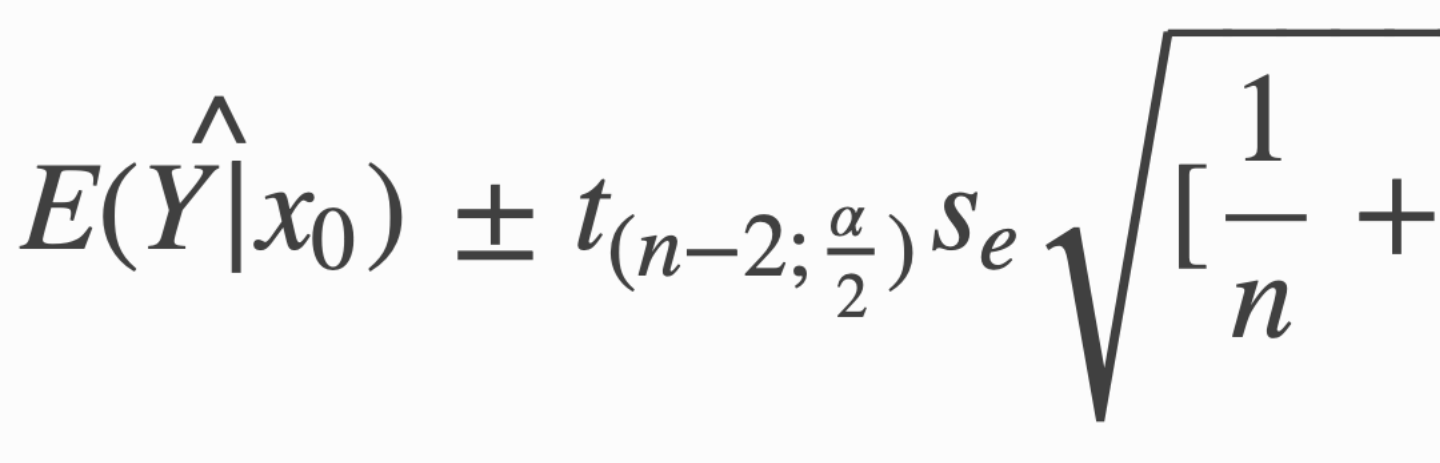
144.9545< β_0 <218.4295

Yang dapat dimaknai bahwa dalam taraf kepercayaan 95%, diyakini bahwa dugaan parameter β_0 berada dalam selang 144.9545 hingga 218.4295

Selang Kepercayaan Amatan

Misalkan ingin menduga nilai rataan (harapan) amatan $X = 50$. Penduga selang kepercayaan 95% bagi rataan (nilai harapan) tersebut adalah:

```
knitr::include_graphics("/Users/user/Downloads/rataan amatan.png")
```



```
dugaan.amatan <- data.frame(x=50)
dugaan.amatan
```

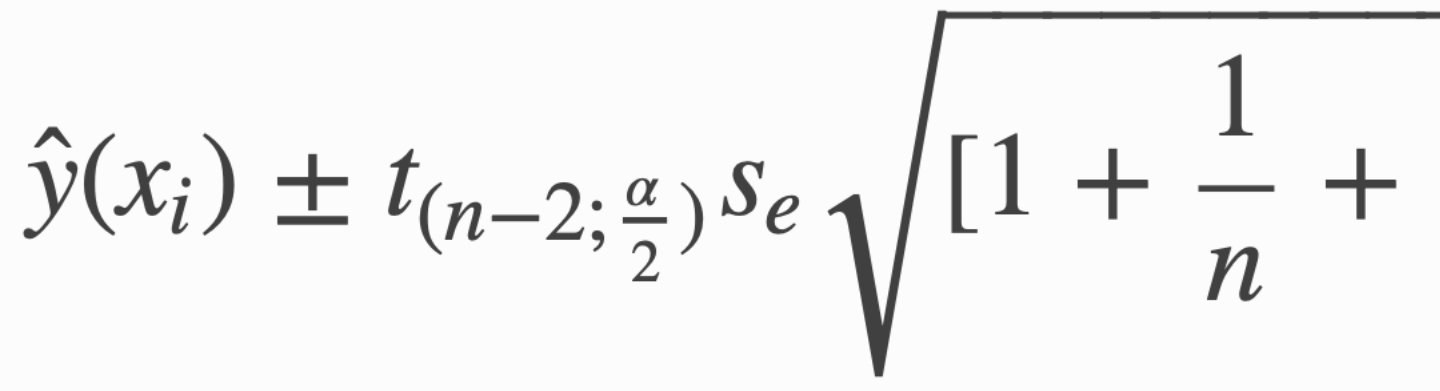
```
##    x
## 1 50
```

```
predict(model, dugaan.amatan, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 244.1237 237.3899 250.8574
```

Penduga selang kepercayaan 95% bagi Individu amatan $X = 50$:

```
knitr::include_graphics("/Users/user/Downloads/individu amatan.png")
```



```r
predict(model, dugaan.amatan, interval = "prediction")
```

```
##        fit      lwr      upr
## 1 244.1237 144.4427 343.8046
```

Sehingga dapat disusun suatu selang kepercayaan 95% untuk $E(\hat{Y}|x_0)$ dan $\hat{y}(x_i)$ dan nilai $X = 50$ sebagai berikut:

237.3899< $E(\hat{Y}|x_0)$ <250.8574

Yang dapat dimaknai bahwa dalam taraf kepercayaan 95%, diyakini bahwa dugaan nilai rataan (harapan) amatan $X = 50$ berada dalam selang 237.3899 hingga 250.8574

144.4427< $\hat{y}(x_i)$ <343.8046

Yang dapat dimaknai bahwa dalam taraf kepercayaan 95%, diyakini bahwa dugaan nilai individu amatan $X = 50$ berada dalam selang 144.4427 hingga <343.804

Tabel Sidik Ragam

```r
dbr <- 1
dbg <- n-2
dbt <- n-1
JKR <- 34799
JKG <- 683825
JKT <- JKR+JKG
KTR <- JKR/dbr
```

```r
SK <- c("Regresi", "Galat", "Total")
db <- c(dbr, dbg, dbt)
JK <- c(JKR, JKG, JKT)
KT <- c(KTR, KTG, NA)
data_frame <- data.frame(SK, db, JK, KT)
data_frame
```

```
##        SK  db     JK        KT
## 1 Regresi   1  34799 34799.000
## 2   Galat 268 683825  2551.587
## 3   Total 269 718624        NA
```

Kesimpulan

Berdasarkan analisis yang kami lakukan, terdapat hubungan positif antara usia dan kadar kolesterol dalam tubuh, yang mengindikasikan bahwa semakin seseorang bertambah usia, maka kemungkinan tinggi juga kadar kolesterolnya. Hal ini menunjukkan adanya risiko yang lebih besar terhadap penyakit jantung. Namun, penting untuk dicatat bahwa faktor usia hanya menjelaskan sebagian kecil dari variasi dalam kadar kolesterol seseorang. Faktor-faktor lain seperti pola makan, aktivitas fisik, dan faktor gaya hidup lainnya juga berperan penting. Oleh karena itu, solusi yang kami rekomendasikan dari hasil analisis ini adalah untuk menjaga kesehatan tubuh dengan menerapkan pola hidup sehat, termasuk pola makan seimbang, rutin berolahraga, mengelola stres dengan baik, menghindari merokok, serta menghindari konsumsi alkohol.