Multiple Linear Regression: LDL and Risk Factors

Statisticians' World

December 4, 2024

Introduction

In this project, we perform a multiple linear regression analysis to understand how several predictors — tobacco use, obesity, alcohol consumption, and age — influence LDL cholesterol levels.

Load Required Packages

```
# Install (if not already installed) - Run interactively only
# install.packages(c("tidyverse", "gtsummary", "ggplot2", "ggpubr",
#                    "GGally", "rsq", "broom", "broom.helpers", "labelled"))

library(tidyverse)
library(gtsummary)
library(ggplot2)
library(ggpubr)
library(GGally)
library(rsq)
library(broom)
library(broom.helpers)
library(labelled)
```

Import Dataset

```
# Set your working directory appropriately (edit as needed)
# setwd("C:/Users/O&1/OneDrive/Documents/R-Youtube")
data <- read.csv("CHDdata.csv")

# Data preview
dim(data)
```

```
## [1] 462  10
```

```
head(data)
```

```
##   sbp tobacco  ldl adiposity famhist typea obesity alcohol age chd
## 1 160   12.00 5.73     23.11 Present    49   25.30   97.20  52   1
## 2 144    0.01 4.41     28.61  Absent    55   28.87    2.06  63   1
## 3 118    0.08 3.48     32.28 Present    52   29.14    3.81  46   0
## 4 170    7.50 6.41     38.03 Present    51   31.99   24.26  58   1
## 5 134   13.60 3.50     27.78 Present    60   25.99   57.34  49   1
## 6 132    6.20 6.47     36.21 Present    62   30.77   14.14  45   0
```

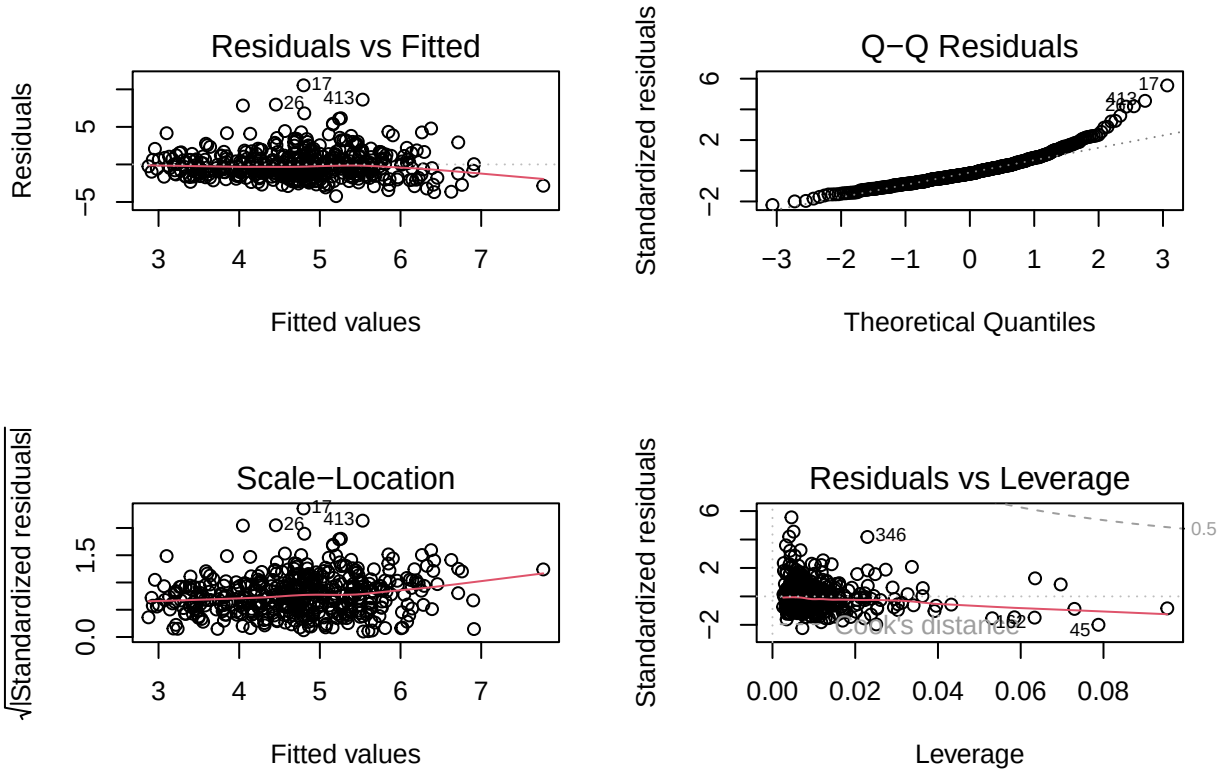
Fit the Multiple Linear Regression Model

```
mlr <- lm(ldl ~ tobacco + obesity + alcohol + age, data = data)
```

5. Assumption Checking

- Independence
  - Assumed, as each patient record is considered independent.
- Linearity, Normality, Homoscedasticity, and Outliers

```
# Show 2x2 diagnostic plots
par(mfrow = c(2, 2))
plot(mlr)
```




```
These plots help us evaluate:

  - Linearity (Residuals vs Fitted)
  - Normality (Q-Q Plot)
  - Homoscedasticity (Scale-Location)
  - Outliers (Residuals vs Leverage)
```

6. Model Summary & Inference

```
summary(mlr)
```

```
##
## Call:
## lm(formula = ldl ~ tobacco + obesity + alcohol + age, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2215 -1.2359 -0.3413  0.8106 10.5324
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.045080   0.566582   0.080   0.9366
## tobacco      0.018324   0.021908   0.836   0.4034
## obesity      0.129752   0.021955   5.910 6.70e-09 ***
## alcohol     -0.006588   0.003690  -1.785   0.0748 .
## age          0.031802   0.007035   4.521 7.87e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.899 on 457 degrees of freedom
## Multiple R-squared:  0.1663, Adjusted R-squared:  0.159
## F-statistic: 22.78 on 4 and 457 DF,  p-value: < 2.2e-16
```

```
tidy(mlr, conf.int = TRUE)
```

```
## # A tibble: 5 x 7
##   term        estimate std.error statistic       p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>         <dbl>    <dbl>     <dbl>
## 1 (Intercept)  0.0451    0.567      0.0796 0.937          -1.07    1.16
## 2 tobacco      0.0183    0.0219     0.836  0.403          -0.0247  0.0614
## 3 obesity      0.130     0.0220     5.91   0.00000000670   0.0866  0.173
## 4 alcohol     -0.00659   0.00369   -1.79   0.0748         -0.0138  0.000663
## 5 age          0.0318    0.00703    4.52   0.00000787      0.0180  0.0456
```

R-Squared

```
rsq(mlr)
```

```
## [1] 0.1662649
```

Summary Table

```
tbl_regression(mlr)
```

Residual Standard Error as % of Mean LDL

```
sigma(mlr) / mean(data$ldl)
```

```
## [1] 0.4006447
```

Characteristic	Beta	95% CI[1]	p-value
tobacco	0.02	-0.02, 0.06	0.4
obesity	0.13	0.09, 0.17	<0.001
alcohol	-0.01	-0.01, 0.00	0.075
age	0.03	0.02, 0.05	<0.001

[1]CI = Confidence Interval

Conclusion

This multiple regression model provides insight into how lifestyle and demographic factors relate to LDL cholesterol. Further diagnostics and interaction modeling could enhance interpretation in future work.

Follow along on YouTube: Statisticians' World