exam project

Mechou lina, Saci nada

2026-01-16

Introduction

This document presents a statistical analysis of cardiovascular data, exploring relationships between physical activity, demographic characteristics, and various health indicators.

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE, fig.align = 'center')

library(ggplot2)
library(ggpubr)
library(car)
library(lmtest)
library(rcompanion)
```

Data Loading

```
dataset01_body_cardio <- read.csv("dataset01_body_cardio.csv")
dataset01<- dataset01_body_cardio
summary(dataset01)
```

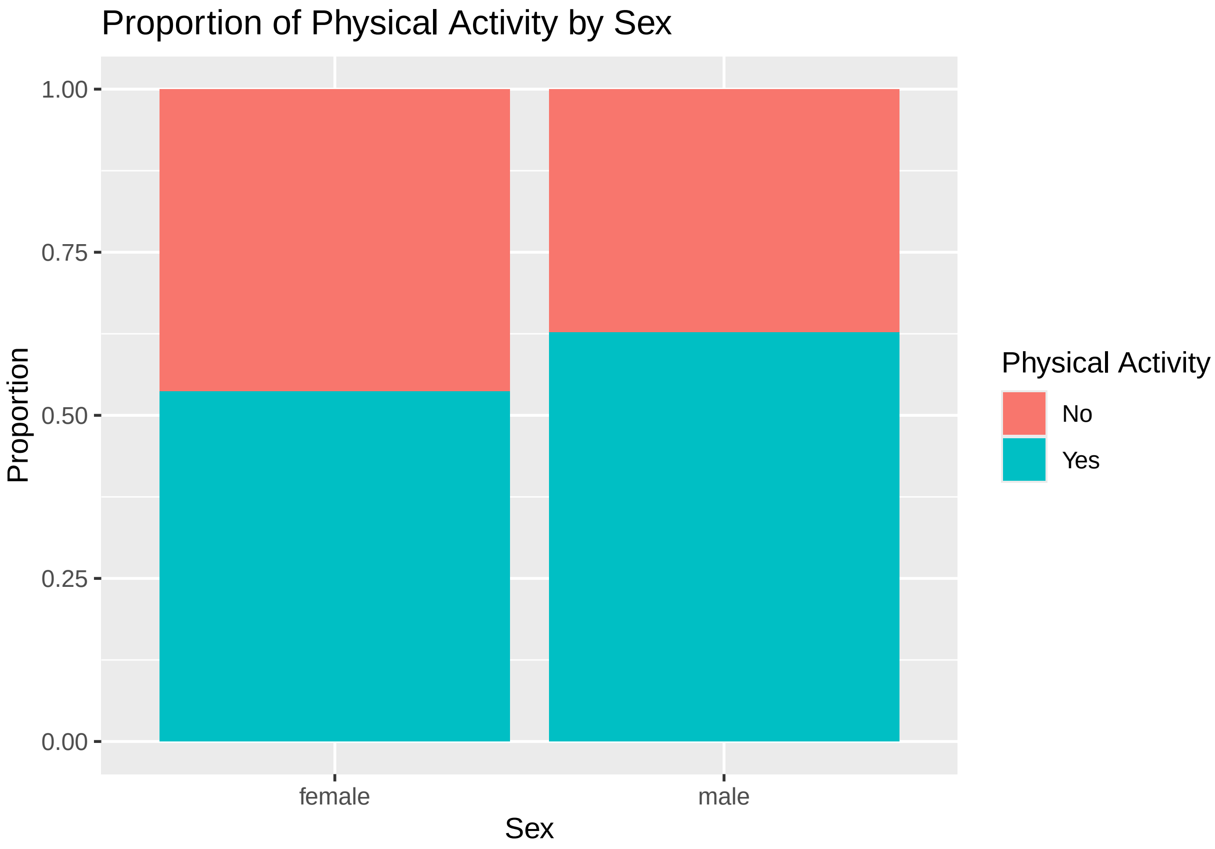
```
##    age_years         sex              height_cm       weight_kg
##  Min.   :12.00   Length:1000        Min.   :141.2   Min.   : 30.00
##  1st Qu.:26.75   Class :character   1st Qu.:161.5   1st Qu.: 65.50
##  Median :42.00   Mode  :character   Median :169.0   Median : 78.90
##  Mean   :42.53                      Mean   :168.8   Mean   : 80.25
##  3rd Qu.:56.00                      3rd Qu.:175.5   3rd Qu.: 91.90
##  Max.   :80.00                      Max.   :199.4   Max.   :223.00
##  body_mass_index systolic_pressure diastolic_pressure   heart_rate
##  Min.   :14.39   Min.   : 82.0     Min.   :  0.00     Min.   : 40.00
##  1st Qu.:23.30   1st Qu.:108.0     1st Qu.: 62.00     1st Qu.: 64.00
##  Median :27.34   Median :118.0     Median : 70.00     Median : 73.00
##  Mean   :28.03   Mean   :119.7     Mean   : 68.98     Mean   : 73.23
##  3rd Qu.:31.69   3rd Qu.:128.0     3rd Qu.: 77.00     3rd Qu.: 80.00
##  Max.   :80.60   Max.   :217.0     Max.   :110.00     Max.   :124.00
##  total_cholesterol direct_cholesterol physically_active
##  Min.   :2.400     Min.   :0.570      Length:1000
##  1st Qu.:4.183     1st Qu.:1.060      Class :character
##  Median :4.860     Median :1.290      Mode  :character
##  Mean   :4.943     Mean   :1.358
##  3rd Qu.:5.590     3rd Qu.:1.600
##  Max.   :9.900     Max.   :3.590
```

1. Physical Activity and Sex

Research Question

Is physical activity level associated with sex?

```
ggplot(data = dataset01, aes(x = sex, fill = physically_active)) +
  geom_bar(position = "fill") +
  labs(x = "Sex", y = "Proportion", fill = "Physical Activity",title = "Proportion of Physical Activity
```



Statistical Test

```
# H0: Physical activity level is independent of sex
# H1: Physical activity level is associated with sex
chisq.test(dataset01$sex, dataset01$physically_active)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dataset01$sex and dataset01$physically_active
## X-squared = 8.0835, df = 1, p-value = 0.004467
```

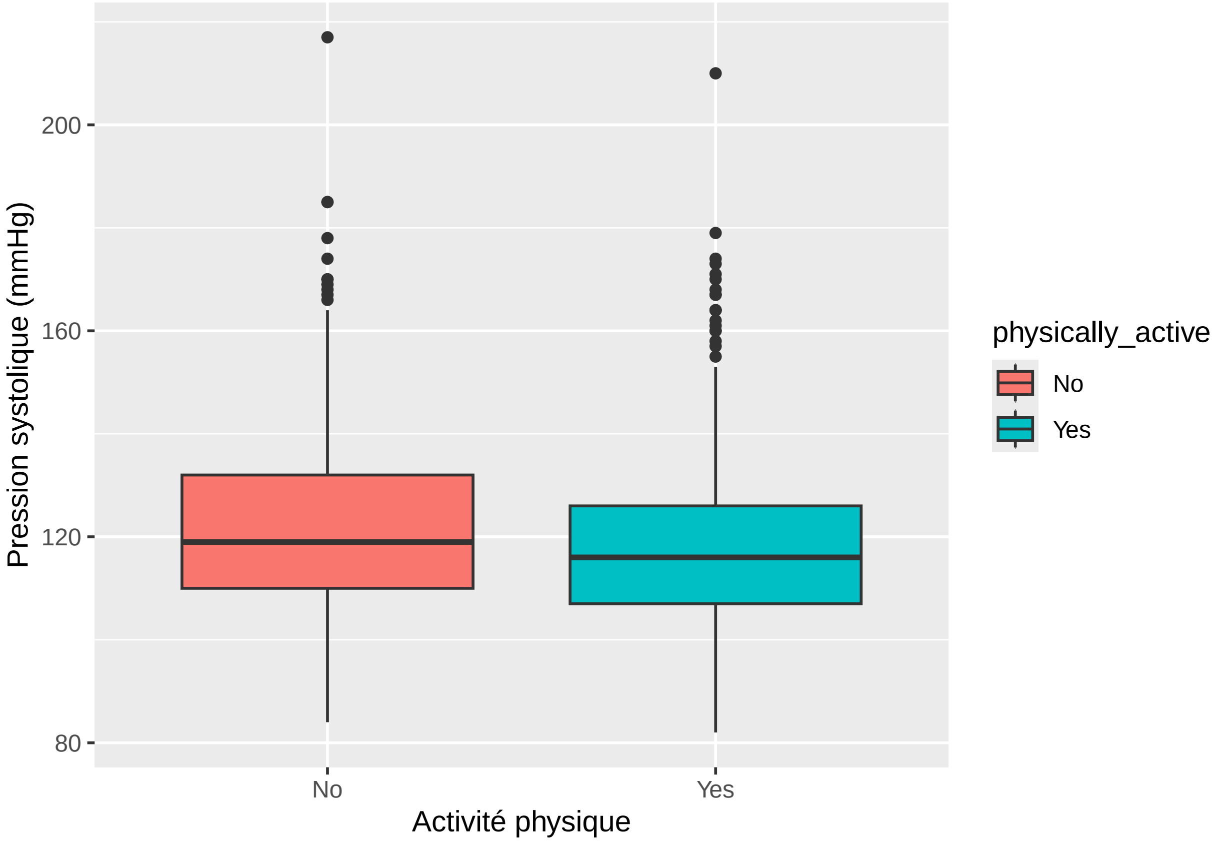
Conclusion: p-value < 0.05, we reject H0. Physical activity level is associated with sex.

2. Systolic Pressure and Physical Activity

Research Question

Does mean systolic blood pressure differ between active and inactive individuals?

```
ggplot(dataset01, aes(x=physically_active, y=systolic_pressure, fill=physically_active))+
  geom_boxplot()+ labs(x="Activité physique", y="Pression systolique (mmHg)")
```



Statistical Tests

```
Active<-dataset01$systolic_pressure[dataset01$physically_active == "Yes"]
Not_active<-dataset01$systolic_pressure[dataset01$physically_active == "No"]
```

Normality Test

```
# H0: Data follows a normal distribution
# H1: Data does not follow a normal distribution
shapiro.test(Active)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Active
## W = 0.94944, p-value = 3.249e-13
```

```
shapiro.test(Not_active)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Not_active
## W = 0.95083, p-value = 1.432e-10
```

Result: p-value $<$ 0.05 for both groups → data does not follow a normal distribution.

Wilcoxon Test

```
# H0: Mean systolic blood pressure is the same in both groups
# H1: Mean systolic blood pressure differs between the two groups
wilcox.test(Active, Not_active)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Active and Not_active
## W = 105441, p-value = 0.0003228
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion: p-value $<$ 0.05, we reject H0. Mean systolic blood pressure is different between active and inactive individuals.

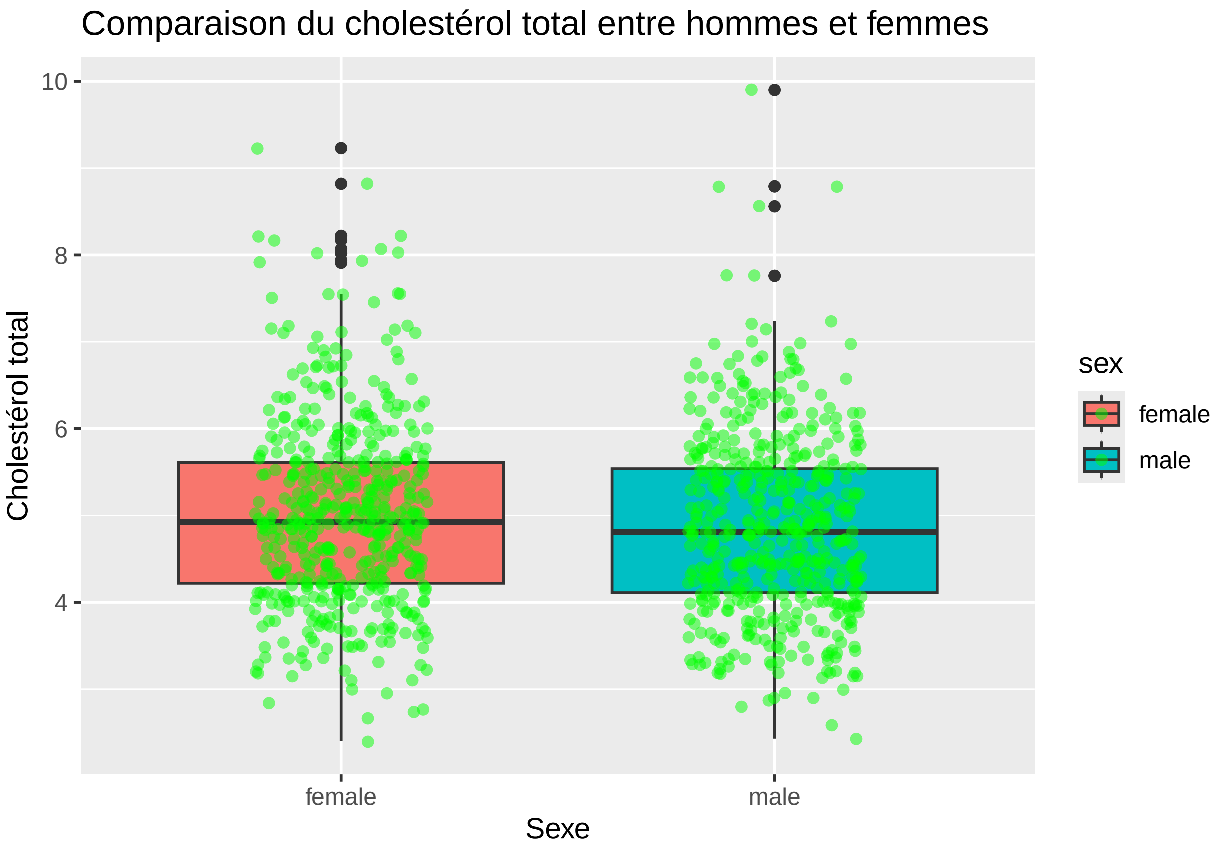
---

3. Total Cholesterol and Sex

Research Question

Do women have a different mean total cholesterol than men?

```
ggplot(dataset01, aes(x = sex, y = total_cholesterol, fill = sex)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.5, color = "green") + labs(x = "Sexe",y = "Cholestérol total",titl
```



Statistical Tests

```
femme<-dataset01$total_cholesterol[dataset01$sex=="female"]
homme<-dataset01$total_cholesterol[dataset01$sex=="male"]
```

Normality Test

```
shapiro.test(femme)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  femme
## W = 0.97766, p-value = 6.867e-07
```

```
shapiro.test(homme)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  homme
## W = 0.97638, p-value = 2.849e-07
```

Result: p-value < 0.05 for both variables (female and male), we reject H0 and accept H1. The data does not follow a normal distribution.

Wilcoxon Test

```
# H0: Mean total cholesterol is the same in women and men
# H1: Women have a different mean total cholesterol than men
wilcox.test(femme, homme)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  femme and homme
## W = 134326, p-value = 0.04096
## alternative hypothesis: true location shift is not equal to 0
```

Conclusion: p-value < 0.05, we reject H0. Women have a different mean total cholesterol than men.

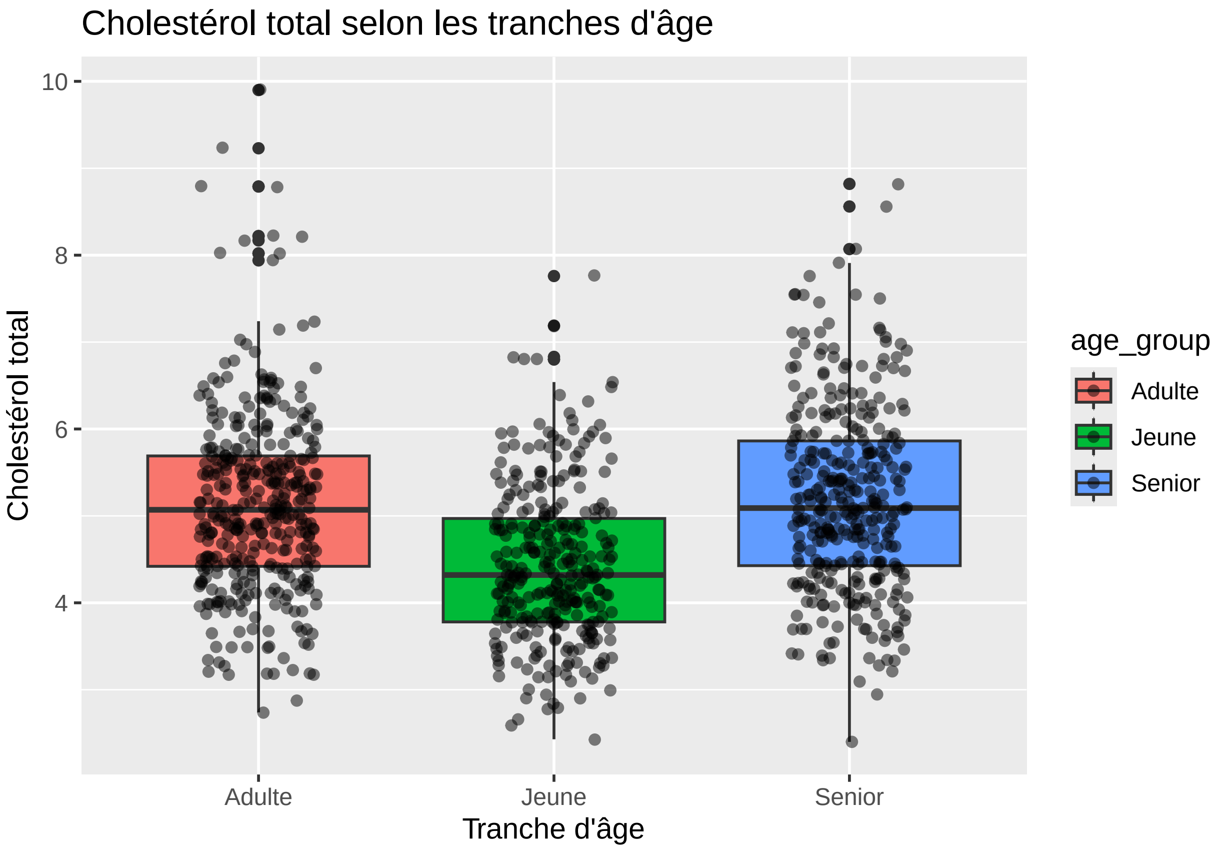
---

4. Total Cholesterol and Age Groups

Research Question

Does mean total cholesterol differ between different age groups (young, adults, seniors)?

```
dataset01$age_group <- ifelse(dataset01$age_years < 30, "Jeune",
                              ifelse(dataset01$age_years <= 50, "Adulte", "Senior"))
ggplot(dataset01, aes(x = age_group, y = total_cholesterol, fill = age_group)) +
  geom_boxplot() +
  geom_jitter(width = 0.2, alpha = 0.5, color = "black") + labs(title = "Cholestérol total selon les tra
```



Statistical Tests

```r
model = lm(total_cholesterol~age_group, data=dataset01)
```

Normality Test of Residuals

```r
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.97988, p-value = 1.564e-10
```

Result: p-value < 0.05, we reject H0 and accept H1. The data does not follow a normal distribution.

Variances Test

```r
# H0: Variances are equal
# H1: Variances are not equal
bartlett.test(total_cholesterol ~ age_group, data = dataset01)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  total_cholesterol by age_group
## Bartlett's K-squared = 11.68, df = 2, p-value = 0.002909
```

Result: p-value < 0.05, we reject H0 and accept H1. The variances are different.

Kruskal-Wallis Test

```r
# H0: Total cholesterol distributions are identical for all age groups
# H1: At least one age group has a different distribution
kruskal.test(total_cholesterol ~ age_group, data = dataset01)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  total_cholesterol by age_group
## Kruskal-Wallis chi-squared = 103.92, df = 2, p-value < 2.2e-16
```

Result: p-value < 0.05, we reject H0 and accept H1. There is a difference between at least two groups.

Post-hoc Comparisons

```r
pairwise.wilcox.test(dataset01$total_cholesterol, dataset01$age_group,p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  dataset01$total_cholesterol and dataset01$age_group
##
##        Adulte Jeune
## Jeune  <2e-16 -
## Senior 1      <2e-16
##
## P value adjustment method: bonferroni
```

Conclusion:

- Young vs Adult: p-value < 0.05 → significant difference
- Young vs Senior: p-value < 0.05 → significant difference

- Adult vs Senior: p-value > 0.05 → no significant difference

Young people have different mean total cholesterol levels than adults and seniors.

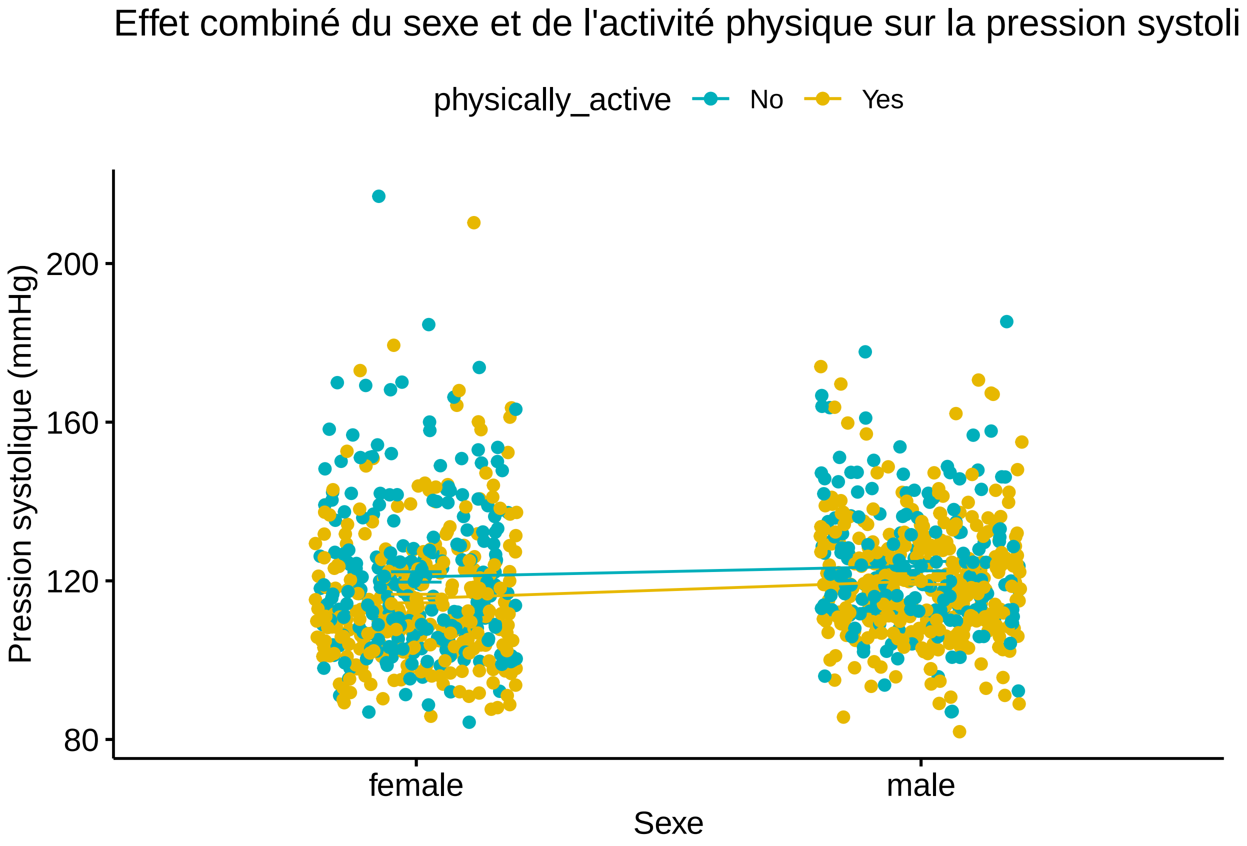
---

5. Combined Effect of Sex and Physical Activity

Research Question

Is there a combined effect of sex and physical activity on systolic pressure?

```
ggline(dataset01, x = "sex",y = "systolic_pressure",color = "physically_active", add = c("mean_se", "ji
       palette = c("#00AFBB", "#E7B800"), ylab = "Pression systolique (mmHg)",xlab = "Sexe",title = "Ef
```



Statistical Tests

```
model=lm(systolic_pressure~sex*physically_active, data=dataset01)
```

Normality Test

```
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.94456, p-value < 2.2e-16
```

Result: p-value < 0.05, we reject H0 and accept H1. The data does not follow a normal distribution.

Variance test

```
# H0: Variances are equal
# H1: At least one group has a different variance
leveneTest(systolic_pressure~sex*physically_active, data=dataset01)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)   
## group   3  5.3359 0.001195 **
##       996                    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Result: p-value < 0.05, we reject H0 and accept H1. The variances between groups are not equal.

Scheirer-Ray-Hare Test

```
# H0: The effect of physical activity on systolic pressure does not depend on sex
# H1: The effect of physical activity on systolic pressure depends on sex
scheirerRayHare(systolic_pressure ~ sex * physically_active, data = dataset01)
```

```
##
## DV:  systolic_pressure
## Observations:  1000
## D:  0.99952
## MS total:  83416.67
```

```
##                         Df   Sum Sq       H p.value
## sex                      1  2236508 26.8242 0.00000
## physically_active        1  1372506 16.4615 0.00005
## sex:physically_active    1    42866  0.5141 0.47336
## Residuals              996 79935503
```

Conclusion: p-value > 0.05, we accept H0. The effect of physical activity is similar in men and women (no significant interaction).

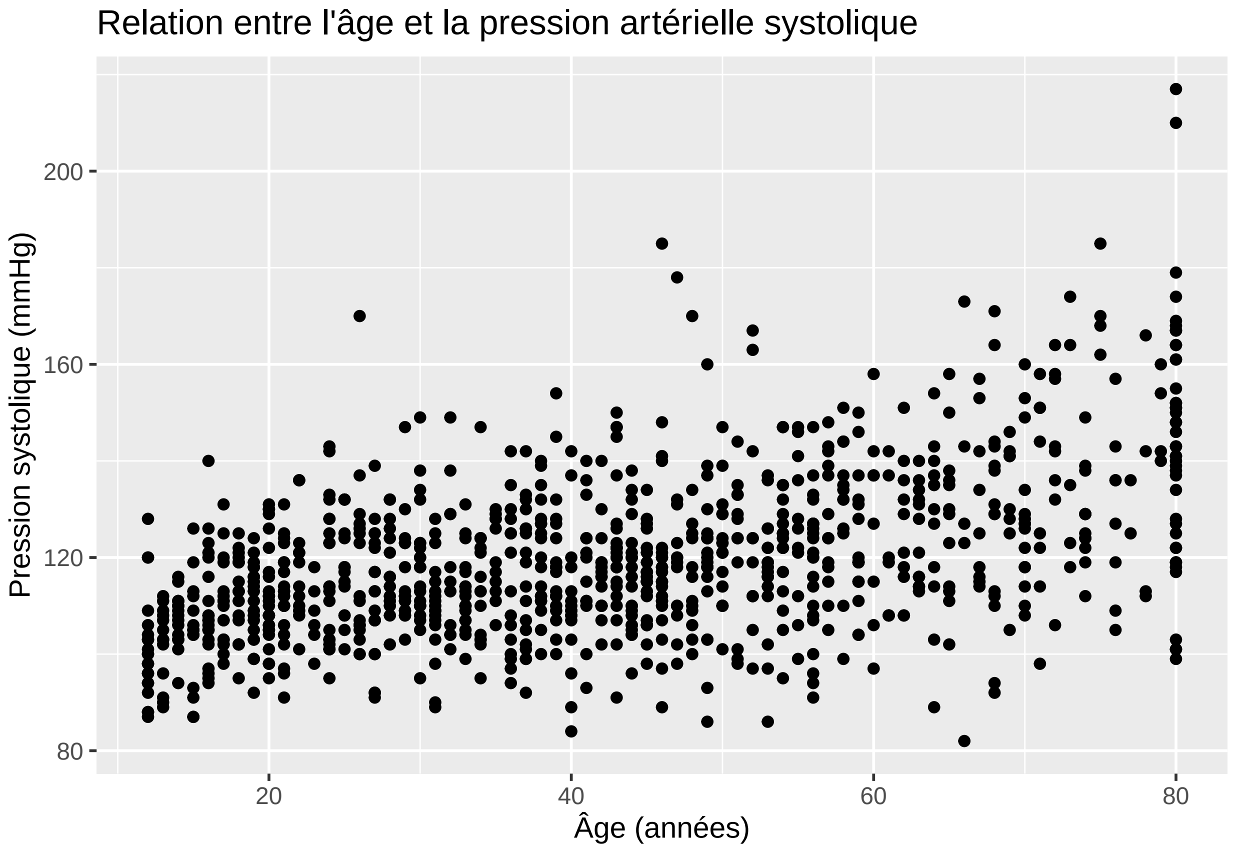
---

6. Correlation Between Age and Systolic Pressure

Research Question

Is age correlated with systolic blood pressure?

```r
ggplot(dataset01, aes(x = age_years, y = systolic_pressure)) +
  geom_point() +labs(title = "Relation entre l'âge et la pression artérielle systolique",x = "Âge (anné
```



Normality Test

```r
shapiro.test(dataset01$age_years)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset01$age_years
## W = 0.96112, p-value = 1.157e-15
```

```r
shapiro.test(dataset01$systolic_pressure)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dataset01$systolic_pressure
## W = 0.94894, p-value < 2.2e-16
```

Result: p-value < 0.05 for both variables (age_years and systolic_pressure), we reject H0 and accept H1. The data does not follow a normal distribution.

Spearman Correlation

```r
# H0: There is no correlation between age and systolic pressure (  = 0)
# H1: There is a correlation between age and systolic pressure (   0)
cor.test(dataset01$age_years, dataset01$systolic_pressure, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  dataset01$age_years and dataset01$systolic_pressure
## S = 87574243, p-value < 2.2e-16
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##      rho
## 0.474554
```

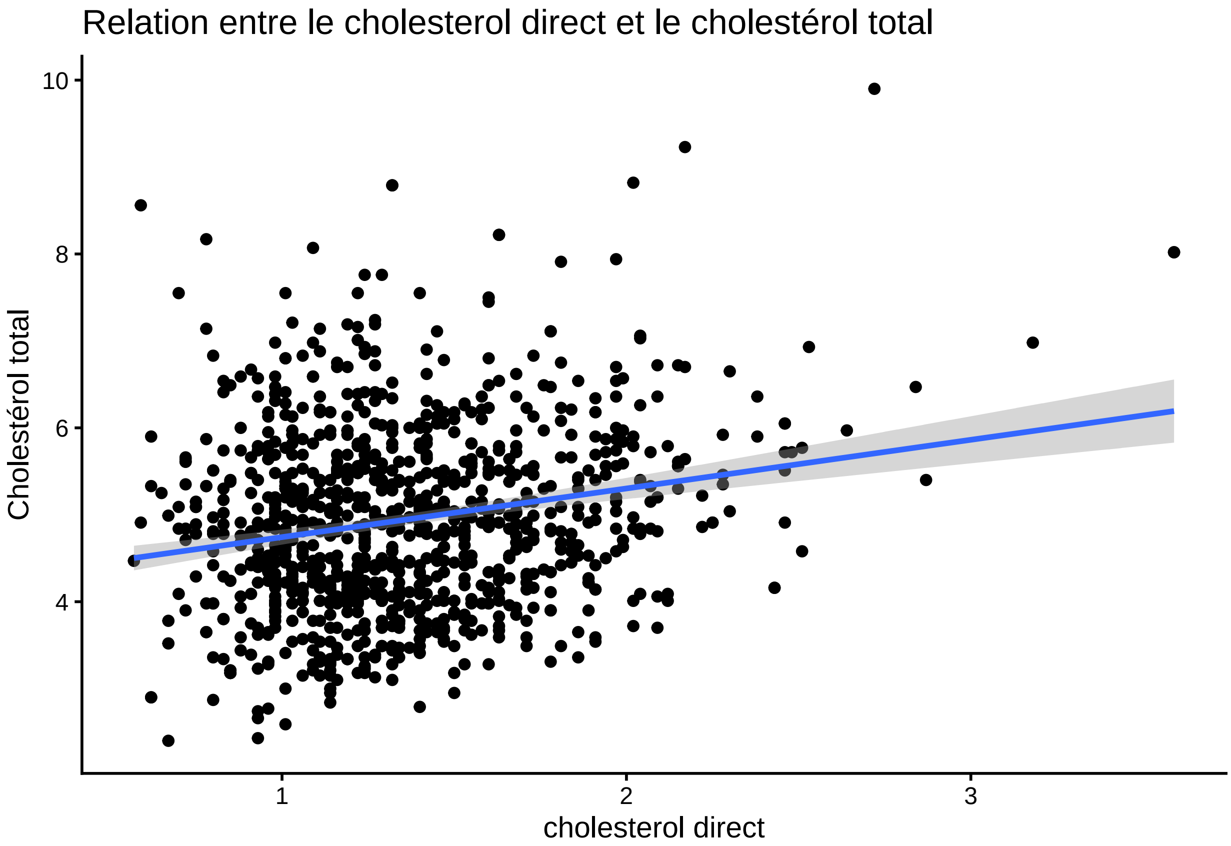
Conclusion: p-value $<$ 0.05, we reject H0. There is a correlation between age and systolic pressure.

7. Regression: Direct and Total Cholesterol

Research Question

Can we predict total cholesterol from direct cholesterol?

```r
ggplot(dataset01, aes(x = direct_cholesterol, y = total_cholesterol)) +
  geom_point() +
  theme_classic() + labs(x = "cholesterol direct",y = "Cholestérol total",title = "Relation entre le cho
  geom_smooth(method = lm)
```



Statistical Tests

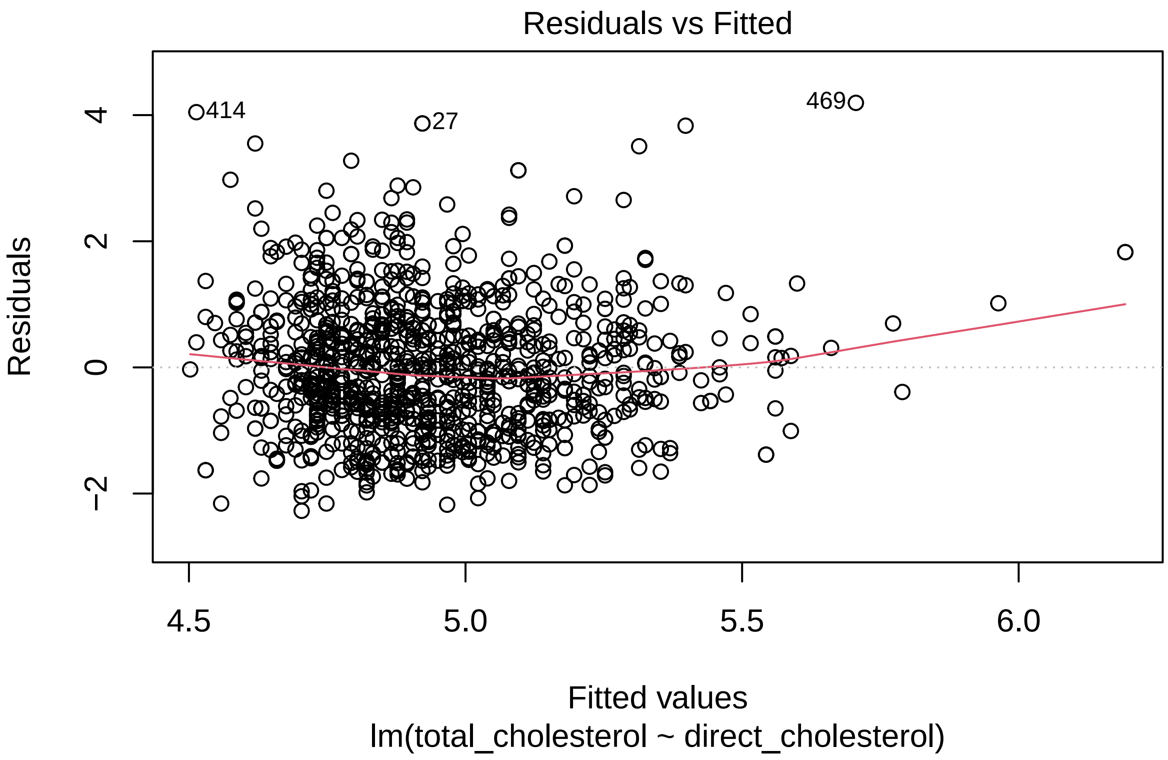
```
# H0: Direct cholesterol has no effect on total cholesterol (  = 0)
# H1: Direct cholesterol has an effect on total cholesterol (    0)
model = lm(total_cholesterol~ direct_cholesterol, dataset01)
summary(model)
```

```
##
## Call:
## lm(formula = total_cholesterol ~ direct_cholesterol, data = dataset01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.2738 -0.7546 -0.0863  0.6264  4.1943
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.18319    0.11555  36.201  < 2e-16 ***
## direct_cholesterol  0.55976    0.08156   6.864 1.18e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.042 on 998 degrees of freedom
## Multiple R-squared:  0.04508,    Adjusted R-squared:  0.04412
## F-statistic: 47.11 on 1 and 998 DF,  p-value: 1.178e-11
```

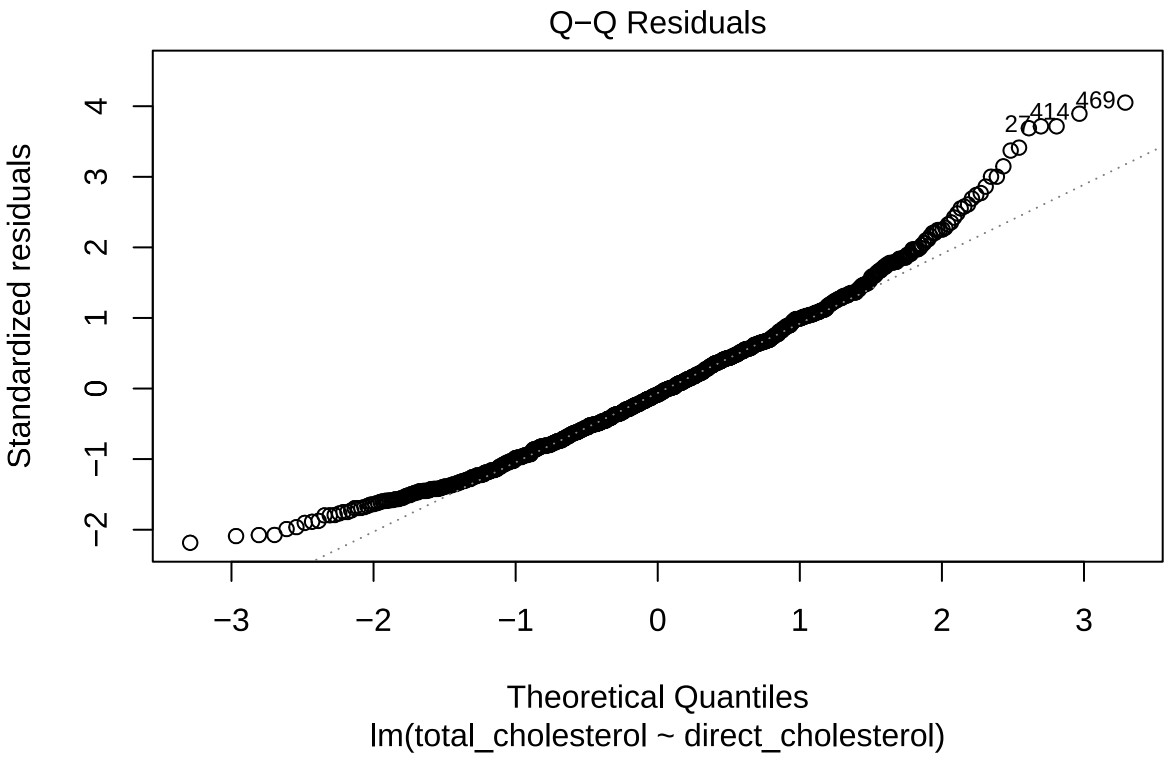
Conclusion: p-value < 0.05, we reject H0. Direct cholesterol has an effect on total cholesterol.

Model Diagnostics

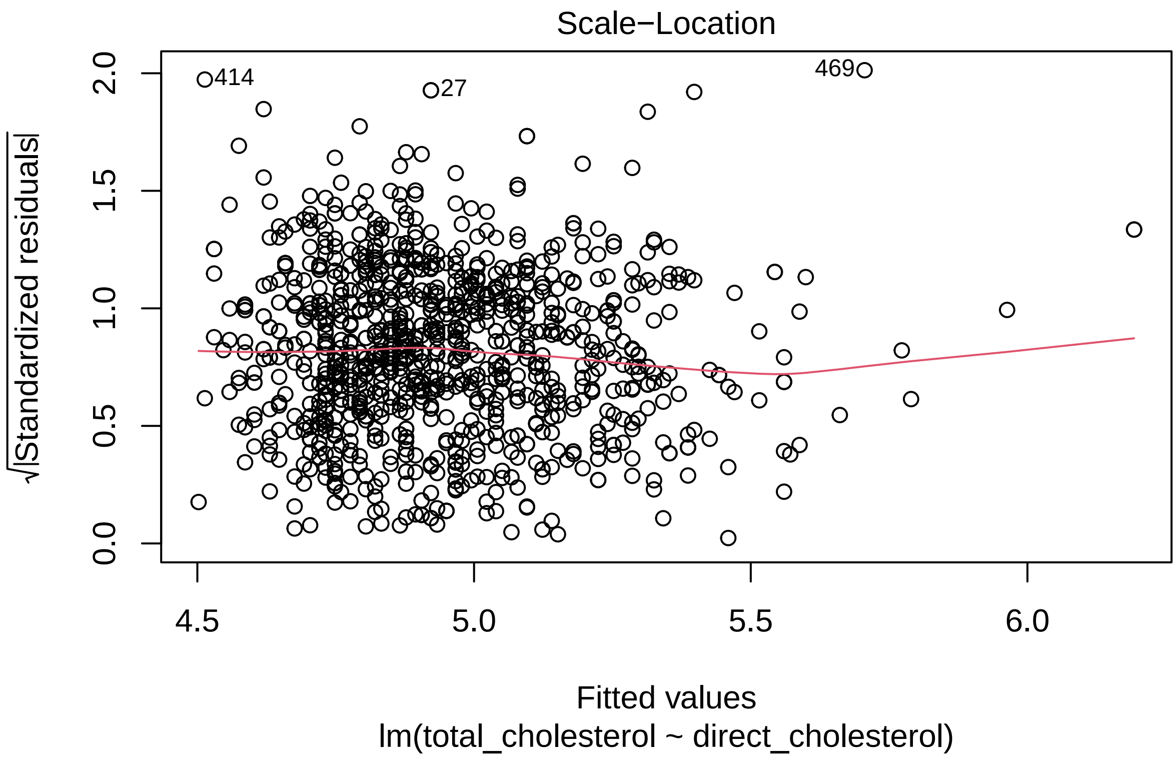
```
plot(model,1)
```



```
plot(model,2)
```



```
plot(model,3)
```



Variance test

```
# H0: Variance of the data is constant
# H1: Variance of the data is not constant
bptest(model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 0.051759, df = 1, p-value = 0.82
```

Result: p-value > 0.05, we accept H0. Variances are constant (homoscedasticity assumption met).

Normality of Residuals

```
# H0: Data follows a normal distribution
# H1: Data does not follow a normal distribution
shapiro.test(model$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  model$residuals
## W = 0.97667, p-value = 1.38e-11
```

Result: p-value < 0.05, we reject H0. The data does not follow a normal distribution.

Summary

This analysis explored multiple relationships in cardiovascular health data:

1. Physical activity is different between men and women.
2. Systolic blood pressure is different in active and inactive people.
3. Cholesterol levels are not the same in women and men.
4. Young people have different cholesterol levels than adults and seniors.
5. Physical activity affects systolic pressure in the same way for men and women.
6. Age is related to systolic blood pressure.
7. Direct cholesterol is linked to total cholesterol.