heart-disease-uci

March 25, 2019

```
In [1]: options(digits=3)
```

1 Heart Disease Dataset

Rosa de Haan (haan1700@student.nhl.nl), Lars Rotgers (rotg1700@student.nhl.nl), Toegepaste Wiskunde, 25 maart 2019.

2 Opdracht

Voor de eerste opdracht voor multivariate statistiek zijn er de volgende producteisen:

1. Visuele inspectie met plots.
2. Correlatiematrix en onderzoek collineariteit.
3. Bepaal regressievergelijkingen gebaseerd op alle verklarende variabelen.
4. Mogelijke uitkomsten berekenen en controleren.
5. Bepaal de verklaarde/onverklaarde en de totale variantie.
6. Bepaal de determinatiecoefficient R^2.
7. Bepaal het best passende model met één variabele.
8. Controleer de uitkomst van de regressielijn.
9. Bepaal het best passende model met twee variabelen.
10. Controleer de uitkomst van de regressielijn.
11. Bepaal of het zinvol is om meer variabelen toe te voegen aan het model.
12. Bepaal de regressielijn met de relevante variabelen.

Voor de opdracht maken we gebruik van de *heart disease dataset*. Deze dataset is verkregen via de onderstaande link:

- https://www.kaggle.com/ronitf/heart-disease-uci

Oorspronkelijk is het de bedoeling dat met behulp van machine learning een model wordt opgesteld om vast te stellen of een patiënt hartziekten heeft. Dit wordt aangegeven in de kolom `target`.

3 Gegevens inladen

```
In [2]: df = read.csv('heart.csv')
```

```r
# kolomnaam herstellen; er staat '`..age'
names = colnames(df);
names[1] = 'age'
colnames(df) = names

head(df)
print(paste('Er zijn', nrow(df), 'rijen, en', length(df), ' kolommen.'))
```

age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

```
[1] "Er zijn 303 rijen, en 14  kolommen."
```

4 Beschrijving van de kolommen

In de dataset zitten de volgende kolommen:

1. `age`: leeftijd. (Ratio)
2. `sex`: geslacht. (Nominaal)
3. `cp`: chest pain type (4 values). (Nominaal)
4. `trestbps`: resting blood pressure. (Ratio)
5. `chol`: serum cholestoral in mg/dl. (Ratio)
6. `fbs`: fasting blood sugar > 120 mg/dl. (Nominaal)
7. `restecg`: resting electrocardiographic results (values 0, 1, 2). (Nominaal)
8. `thalanch`: maximum heartrate achieved. (Ratio)
9. `exang`: exercise induced angina. (Nominaal)
10. `oldpeak`: ST depression induced by exercise relative to test. (Ratio)
11. `slope`: the slope of the peak exercise ST segment. (Nominaal)
12. `ca`: number of major vessels (0-3) color by flourosopy. (Ordinaal)
13. `thal`: 3 = normal, 6 = fixed defect, 7 = reversable defect. (Nominaal)
14. `target`: indicated if someone has heart disease, 0 = false, 1 = true. (Nominaal)

5 Beschrijvende statistiek

Om een eerste indruk te krijgen van de gegevens in de dataset bepalen we met `summary` enkele algemene gegevens:

```
In [3]: summary(df)
```

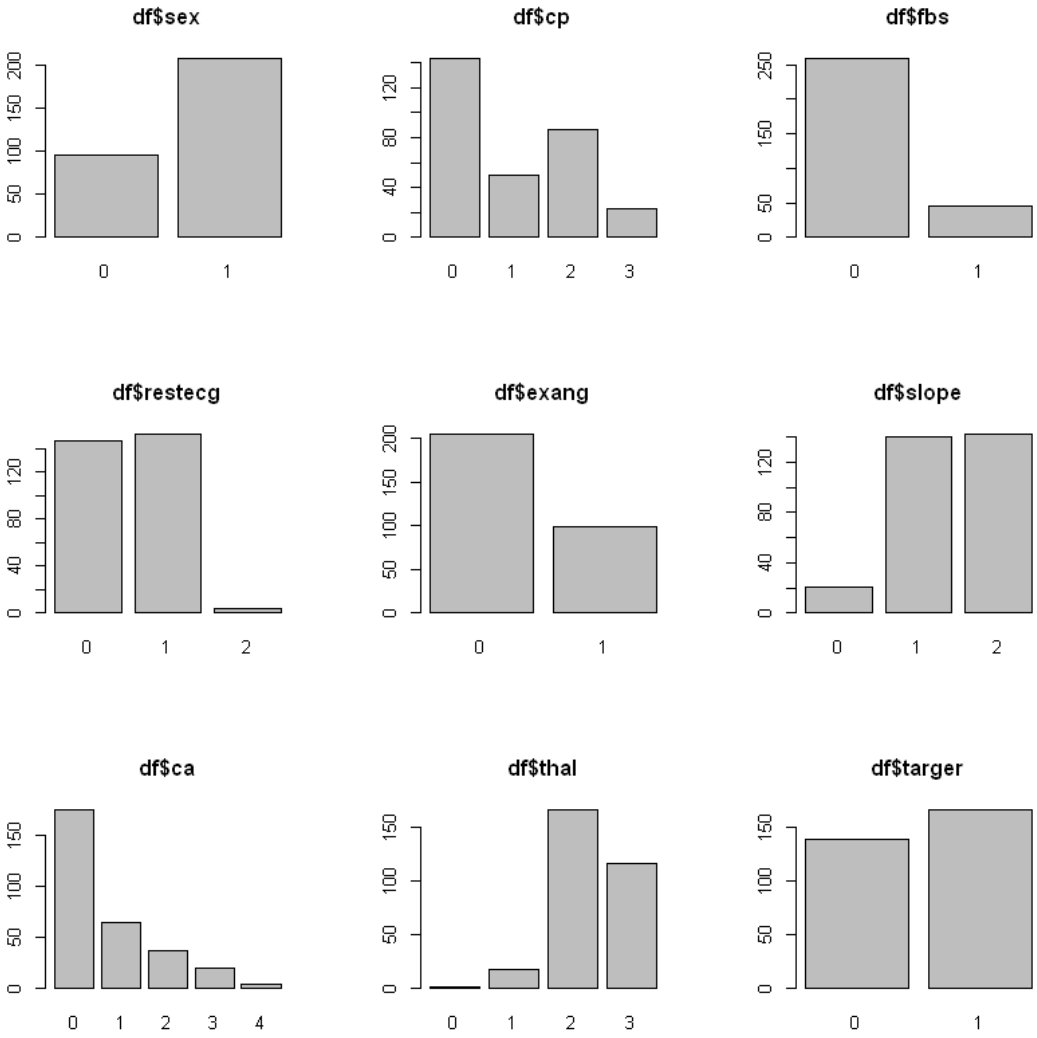
```
      age             sex               cp           trestbps         chol
 Min.   :29.0   Min.   :0.000   Min.   :0.000   Min.   : 94   Min.   :126
 1st Qu.:47.5   1st Qu.:0.000   1st Qu.:0.000   1st Qu.:120   1st Qu.:211
```

```
 Median :55.0   Median :1.000   Median :1.000   Median :130   Median :240
 Mean   :54.4   Mean   :0.683   Mean   :0.967   Mean   :132   Mean   :246
 3rd Qu.:61.0   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:140   3rd Qu.:274
 Max.   :77.0   Max.   :1.000   Max.   :3.000   Max.   :200   Max.   :564
      fbs            restecg         thalach        exang          oldpeak
 Min.   :0.000   Min.   :0.000   Min.   : 71   Min.   :0.000   Min.   :0.00
 1st Qu.:0.000   1st Qu.:0.000   1st Qu.:134   1st Qu.:0.000   1st Qu.:0.00
 Median :0.000   Median :1.000   Median :153   Median :0.000   Median :0.80
 Mean   :0.149   Mean   :0.528   Mean   :150   Mean   :0.327   Mean   :1.04
 3rd Qu.:0.000   3rd Qu.:1.000   3rd Qu.:166   3rd Qu.:1.000   3rd Qu.:1.60
 Max.   :1.000   Max.   :2.000   Max.   :202   Max.   :1.000   Max.   :6.20
     slope            ca            thal           target
 Min.   :0.0   Min.   :0.00   Min.   :0.00   Min.   :0.000
 1st Qu.:1.0   1st Qu.:0.00   1st Qu.:2.00   1st Qu.:0.000
 Median :1.0   Median :0.00   Median :2.00   Median :1.000
 Mean   :1.4   Mean   :0.73   Mean   :2.31   Mean   :0.545
 3rd Qu.:2.0   3rd Qu.:1.00   3rd Qu.:3.00   3rd Qu.:1.000
 Max.   :2.0   Max.   :4.00   Max.   :3.00   Max.   :1.000
```

5.1 Staafdiagrammen voor nominale/ordinale variabelen

Om een idee te krijgen hoe de data is verdeeld binnen de verschillende variabelen worden er staafdiagrammen opgesteld voor de nominale/ordinale variabelen

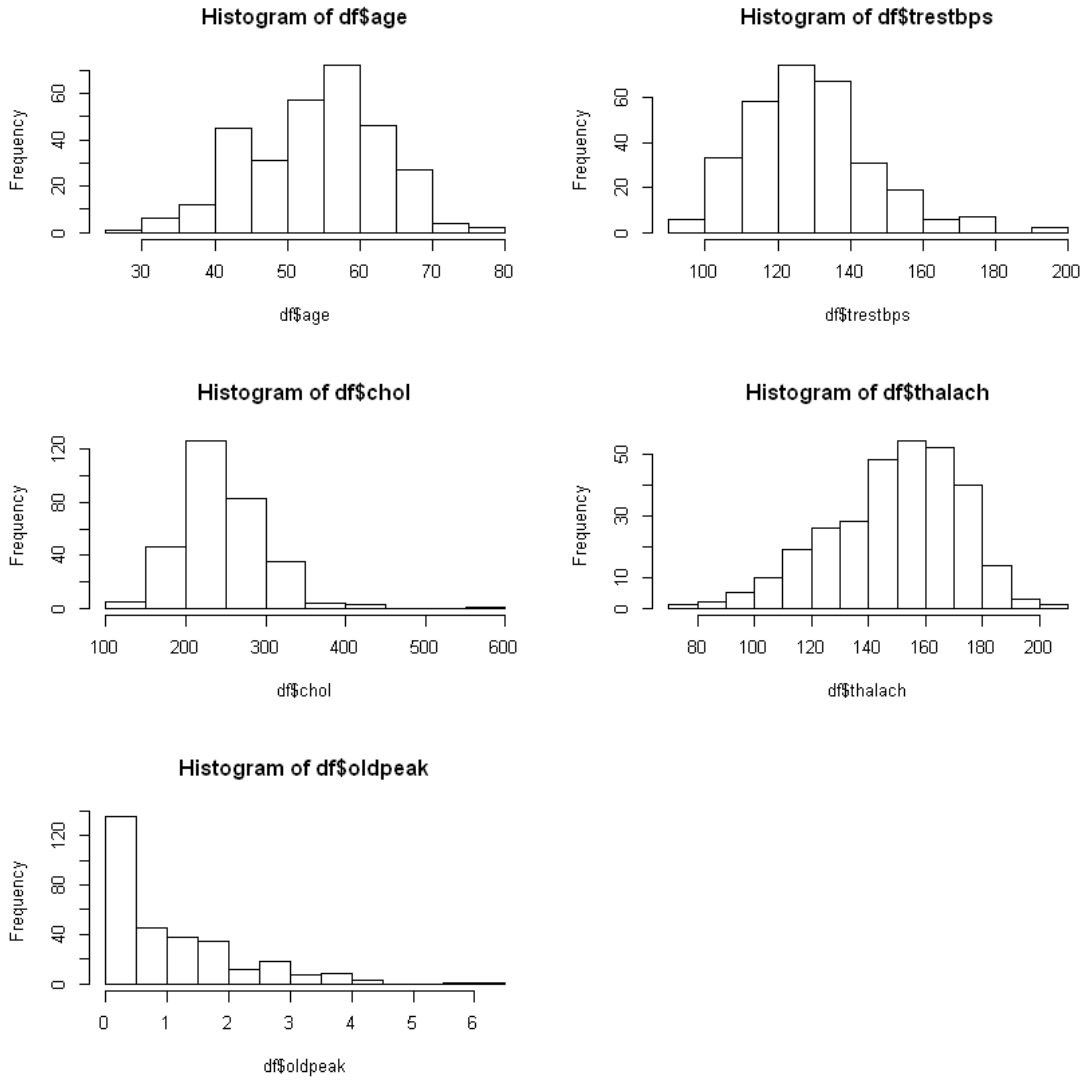
```
In [4]: par(mfrow=c(3,3))
        barplot(table(df$sex), main="df$sex")
        barplot(table(df$cp), main="df$cp")
        barplot(table(df$fbs), main="df$fbs")
        barplot(table(df$restecg), main="df$restecg")
        barplot(table(df$exang), main="df$exang")
        barplot(table(df$slope), main="df$slope")
        barplot(table(df$ca), main="df$ca")
        barplot(table(df$thal), main="df$thal")
        barplot(table(df$target), main="df$targer")
```



5.2 Histogrammen voor ratio variabelen

Hetzelfde doen we ook voor variabelen met het meetniveau ratio, maar dan met histogrammen.

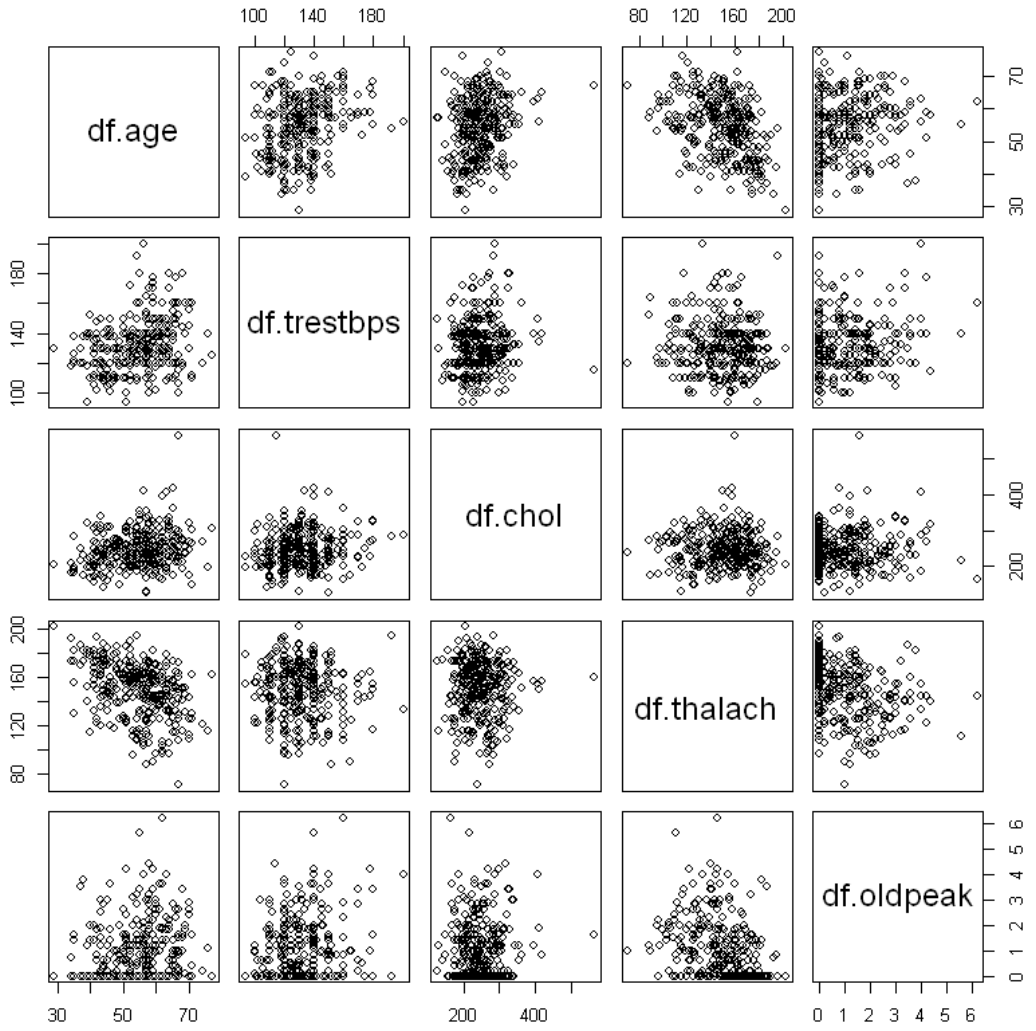
```
In [5]: par(mfrow=c(3, 2))
        hist(df$age)
        hist(df$trestbps)
        hist(df$chol)
        hist(df$thalach)
        hist(df$oldpeak)
```



6 Correlatieplots

Om de correlatie tussen de variabelen te analyseren is een correlatieplot handig. Het volgende figuur is een plot van alle variabelen waarvan het meetniveau ratio is.

```
In [6]: plot(data.frame(df$age, df$trestbps, df$chol, df$thalach, df$oldpeak))
```



Kijkende naar de resultaten, is te zien dat de correlaties vrij zwak zijn. Voor de exacte correlatiecoefficienten kijken we naar de correlatiematrix voor de ratio variabelen:

```
In [7]: cor(data.frame(df$age, df$trestbps, df$chol, df$thalach, df$oldpeak))
```

	df.age	df.trestbps	df.chol	df.thalach	df.oldpeak
df.age	1.000	0.2794	0.21368	-0.39852	0.210
df.trestbps	0.279	1.0000	0.12317	-0.04670	0.193
df.chol	0.214	0.1232	1.00000	-0.00994	0.054
df.thalach	-0.399	-0.0467	-0.00994	1.00000	-0.344
df.oldpeak	0.210	0.1932	0.05395	-0.34419	1.000

De correlatie die er het meeste uitspringt is dat de leeftijd een negatief effect, namelijk `-0.3985`, heeft op de maximaal te behalen hartslag.

7 Correlatiematrix en collineariteit

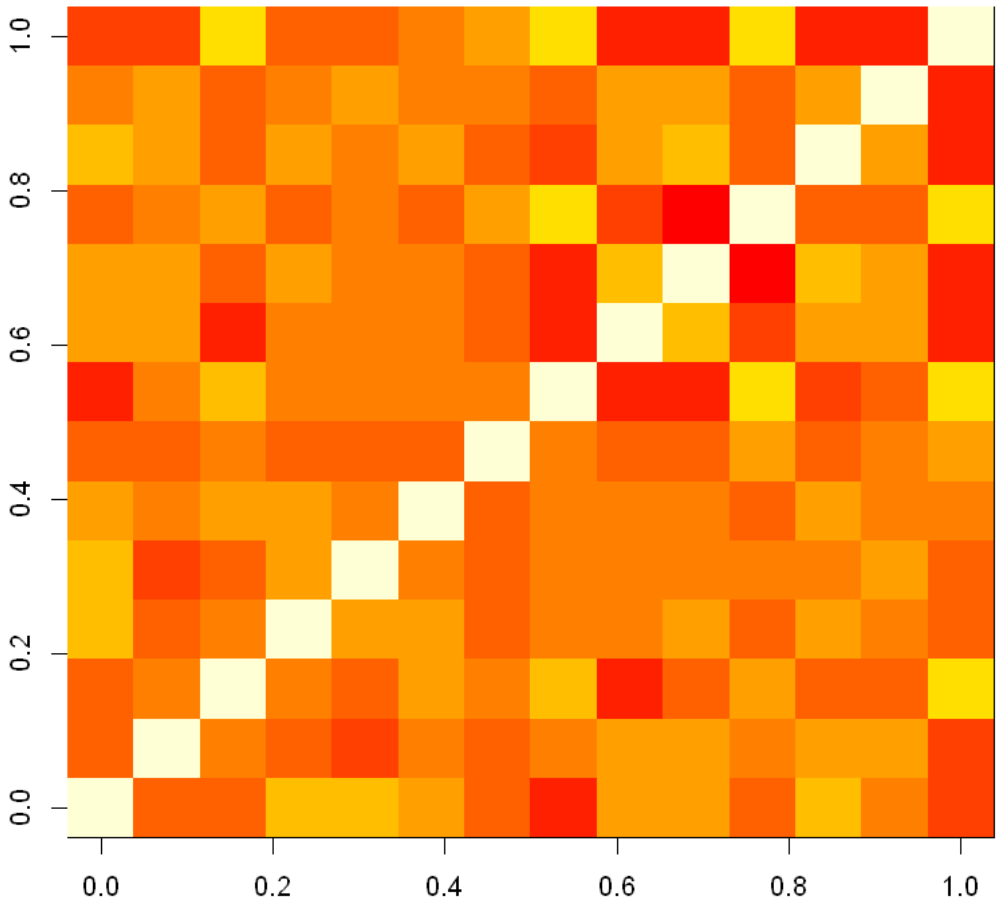
Om uit te zoeken wat de samenhang tussen de variabelen is, is het handig om naar de correlatiematrix te kijken.

```
In [8]: cor(df)
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeal
age	1.0000	-0.0984	-0.0687	0.2794	0.21368	0.12131	-0.1162	-0.39852	0.0968	0.21001
sex	-0.0984	1.0000	-0.0494	-0.0568	-0.19791	0.04503	-0.0582	-0.04402	0.1417	0.09609
cp	-0.0687	-0.0494	1.0000	0.0476	-0.07690	0.09444	0.0444	0.29576	-0.3943	-0.1492
trestbps	0.2794	-0.0568	0.0476	1.0000	0.12317	0.17753	-0.1141	-0.04670	0.0676	0.19322
chol	0.2137	-0.1979	-0.0769	0.1232	1.00000	0.01329	-0.1510	-0.00994	0.0670	0.05395
fbs	0.1213	0.0450	0.0944	0.1775	0.01329	1.00000	-0.0842	-0.00857	0.0257	0.00575
restecg	-0.1162	-0.0582	0.0444	-0.1141	-0.15104	-0.08419	1.0000	0.04412	-0.0707	-0.0587
thalach	-0.3985	-0.0440	0.2958	-0.0467	-0.00994	-0.00857	0.0441	1.00000	-0.3788	-0.3441
exang	0.0968	0.1417	-0.3943	0.0676	0.06702	0.02567	-0.0707	-0.37881	1.0000	0.28822
oldpeak	0.2100	0.0961	-0.1492	0.1932	0.05395	0.00575	-0.0588	-0.34419	0.2882	1.00000
slope	-0.1688	-0.0307	0.1197	-0.1215	-0.00404	-0.05989	0.0930	0.38678	-0.2577	-0.5775
ca	0.2763	0.1183	-0.1811	0.1014	0.07051	0.13798	-0.0720	-0.21318	0.1157	0.22268
thal	0.0680	0.2100	-0.1617	0.0622	0.09880	-0.03202	-0.0120	-0.09644	0.2068	0.21024
target	-0.2254	-0.2809	0.4338	-0.1449	-0.08524	-0.02805	0.1372	0.42174	-0.4368	-0.4307

Ook hier is te zien dat de correlaties tussen de variabelen vrij zwak is. De aanwezigheid van collineariteit is binnen deze dataset niet een probleem. De afwezigheid van enige correlaties echter wel.

```
In [9]: image(cor(df))
```



8 Regressielijnen voor alle verklarende variabelen

Alle verklarende variabelen zijn in dit geval de variabelen die het ratio meetniveau hebben. Met de volgende functie wordt er een regressielijn opgesteld en een grafiek gemaakt met de R^2 waarde.

```
In [10]: lm2 = function(y, x, ylab, xlab) {
             fit = lm(y~x);
             SS.tot = sum((y - mean(y))^2)
             SS.res = sum(fit$residuals^2)
             R2 = 1 - SS.res / SS.tot
             plot(x,y, main=paste("R^2: ", round(R2, 3), sep=""), xlab=xlab, ylab=ylab)
             abline(fit, lw=2, col='red')
```

```
    (fit)
}
```

Vervolgens wordt er voor elke variabele een regressielijn bepaald met bijbehorende R^2 waarde.

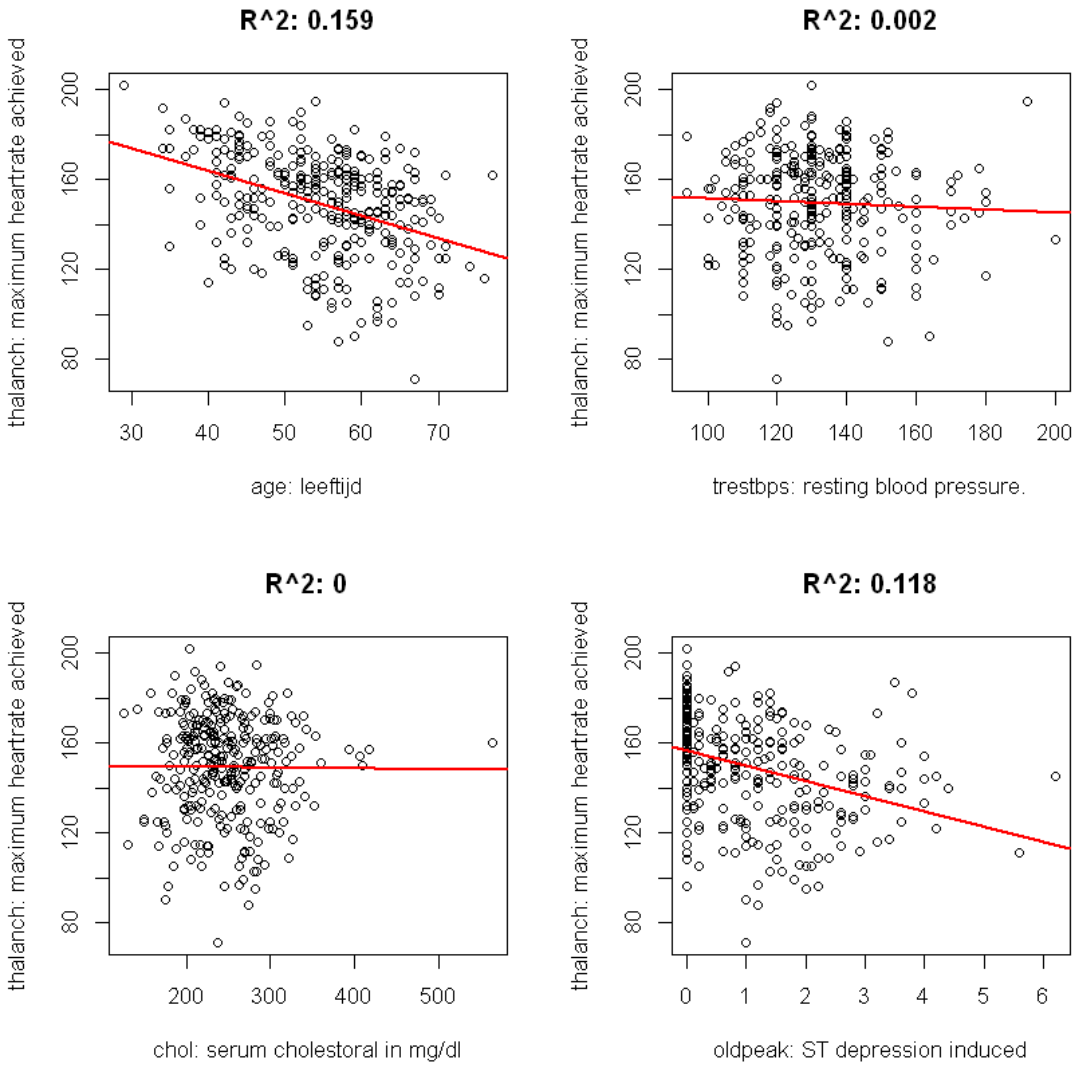
```
In [11]: par(mfrow=c(2, 2))
         lm2(df$thalach, df$age, 'thalanch: maximum heartrate achieved', 'age: leeftijd')
         lm2(df$thalach, df$trestbps, 'thalanch: maximum heartrate achieved', 'trestbps: resti
         lm2(df$thalach, df$chol, 'thalanch: maximum heartrate achieved', 'chol: serum choleste
         lm2(df$thalach, df$oldpeak, 'thalanch: maximum heartrate achieved', 'oldpeak: ST depr
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     204.29        -1.01
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    157.674       -0.061
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
  150.72861     -0.00439
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     156.71        -6.79
```



Het 'beste' model wordt verkregen bij het opstellen van een regressielijn met de leeftijd. Echter geeft een $R^2 = 0.159$ aan dat het model niet echt bruikbaar is. De overige variabelen presteren nog slechter.

Aangezien de dataset oorspronkelijk is bedoeld voor machine learning, lijkt ons het een logische verklaring dat het verkrijgen van een goed model in dit geval niet met een eenvoudige regressielijn gaat lukken.

8.1 Controle regressiemodel

Ter controle voeren we de berekening waarmee we de coefficienten bepalen met de hand uit. De variabele die we gebruiken is die van de leeftijd.

```
In [12]: summary(lm(df$thalach~df$age))
```

```
Call:
lm(formula = df$thalach ~ df$age)

Residuals:
   Min     1Q Median     3Q    Max
-65.95 -11.95   3.97  15.92  44.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  204.289      7.348   27.80  < 2e-16 ***
df$age        -1.005      0.133   -7.54  5.6e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21 on 301 degrees of freedom
Multiple R-squared:  0.159,Adjusted R-squared:  0.156
F-statistic: 56.8 on 1 and 301 DF,  p-value: 5.63e-13
```

Om de coefficienten te vinden, bepalen we:

$$\mathbf{Ax} = \mathbf{b} \implies \mathbf{x} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{b}.$$

```
In [13]: A = cbind(unlist(df$age), 1)
         b = unlist(df$thalach)
         x = solve(t(A) %*% A) %*% t(A) %*% b
         x
```

-1.01
204.29

Om de residuen te vinden, bepalen we:

$$\mathbf{Ax} - \mathbf{b} = \mathbf{e}.$$

```
In [14]: e = A %*% x - b
```

En de R^2 vinden we met:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2}.$$

```
In [15]: SSE = sum(e^2)
         SST = sum((b-mean(b))^2)
         R2 = 1 - SSE / SST
         R2
```

0.158819735163772

9 Meervoudige regressie

Aangezien er maar twee variabelen zijn waar we enigszins iets mee kunnen doen, stellen we hier een model voor op:

```
In [16]: Y = unlist(df$thalach)
         X1 = unlist(df$age)
         X2 = unlist(df$oldpeak)
         fit = lm(Y~X1+X2)
         summary(fit)
```

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
   Min     1Q Median     3Q    Max
-67.99 -11.53   4.01  14.36  39.45

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  202.031      7.056   28.63  < 2e-16 ***
X1            -0.861      0.131   -6.59  2.0e-10 ***
X2            -5.376      1.022   -5.26  2.8e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.2 on 300 degrees of freedom
Multiple R-squared:  0.23,Adjusted R-squared:  0.225
F-statistic: 44.8 on 2 and 300 DF,  p-value: <2e-16
```

Voor dit model is de R^2 waarde 0.2298. Door het combineren van de variabelen is het model toch nog beter geworden.

9.1 Controle meervoudige regressie

Eerst berekenen we alle sommaties van de onderstaande vergelijkingen:

1. $\sum Y_i = nb_0 + b_1 \sum X_{1i} + b_2 \sum X_{2i}$
2. $\sum X_{1i} Y_i = b_0 \sum X_{1i} + b_1 \sum X_{1i}^2 + b_2 \sum X_{1i} X_{2i}$
3. $\sum X_{2i} Y_i = b_0 \sum X_{2i} + b_1 \sum X_{1i} X_{2i} + b_2 \sum X_{2i}^2$

```
In [17]: SY = sum(Y)
         SX1 = sum(X1)
         SX2 = sum(X2)
         SX1Y = sum(X1*Y)
         SX1_2 = sum(X1^2)
```

```
SX1X2 = sum(X1*X2)
SX2Y = sum(X2*Y)
SX2_2 = sum(X2^2)
c(SY, SX1, SX2, SX1Y, SX1_2, SX1X2, SX2Y, SX2_2)
```

1. 45343 2. 16473 3. 315 4. 2440096 5. 920487 6. 17794.2 7. 44374.4 8. 734.6

```
In [18]: length(X1)
```

303

Dit levert de volgende drie lineaire vergelijkingen op:

1. $45343 = 303b_0 + 16473b_1 + 315b_2$
2. $920487 = 16473b_0 + 920487b_1 + 17794.2b_2$
3. $44374.4 = 315b_0 + 17794.2b_1 + 734.6b_2$

Dit stelsel kunnen we eenvoudig oplossen met R:

```
In [19]: A = cbind(c(length(X1), SX1, SX2), c(SX1, SX1_2, SX1X2), c(SX2, SX1X2, SX2_2))
         b = c(SY, SX1Y, SX2Y)
         x = solve(A,b)
         x
```

1. 202.031038398374 2. -0.860739983205589 3. -5.37598378210027

Dit komt overeen met het resultaat van `summary(lm(...))`. Op dezelfde manier als bij enkelvoudige regressie, bepalen we ook hier R^2:

```
In [20]: SSE = sum(fit$residuals^2)
         SST = sum((Y - mean(Y))^2)
         1 - SSE/SST
```

0.229806899136906

10 Conclusie

Door de sterke afwezigheid van correlaties is het moeilijk om een goed model te vinden met behulp van een lineaire regressielijn. Het 'beste' model wat is gevonden heeft een R^2 van 0.23 (afgerond).

In dit geval is lineaire regressie misschien niet de beste optie. Het verkregen model met deze methode is in onze ogen niet bruikbaar. Wellicht levert de vervolgopdacht met logistische regressie betere resultaten.