Opgave 1 Analytics 2019

Naam student: *graag hier invullen*

Data voor de eerste vragen van deze opgave: **crimepubs.csv**. **Zie crimepubs.docx** voor een korte beschrijving van de variabelen.

Voor het tweede deel van de vragen staat de data in **crime2.csv**. Zie **crime2.docx** voor een korte beschrijving van de variabelen.

Inleveren

Inleveren UITSLUITEND via Blackboard. Uiterste datum van inleveren: 3-3-2019.

Inleveren UITSLUITEND in formaat .PDF, .DOC of .DOCX.

Als je met RStudio werkt kan je de .Rmd file als basis gebruiken. Anders mag je de resultaten in dit document opnemen. Vergeet niet je naam in het document te zetten.

Inleiding (vraag 1 t/m3)

De data bevatten een aantal oude gegevens over misdadigheid. De vraag die we gaan behandelen is: Welke factor (of factoren) dragen er toe bij dat men wel of niet misdaden pleegt?

**Vraag 1 (crimepubs.csv)**Inlezen en bewerken van de data.

Eerst moeten we de data inlezen en nutteloze variabelen verwijderen.

**1a)** Laad de data (crimepubs.csv).

**Gevraagde antwoorden:**

Geef alle gebruikte R-commando’s.

> setwd("D:/GitHub/School/P2.3/Analytics/opgaven\_data")

> data=read.csv("crimepubs.csv")

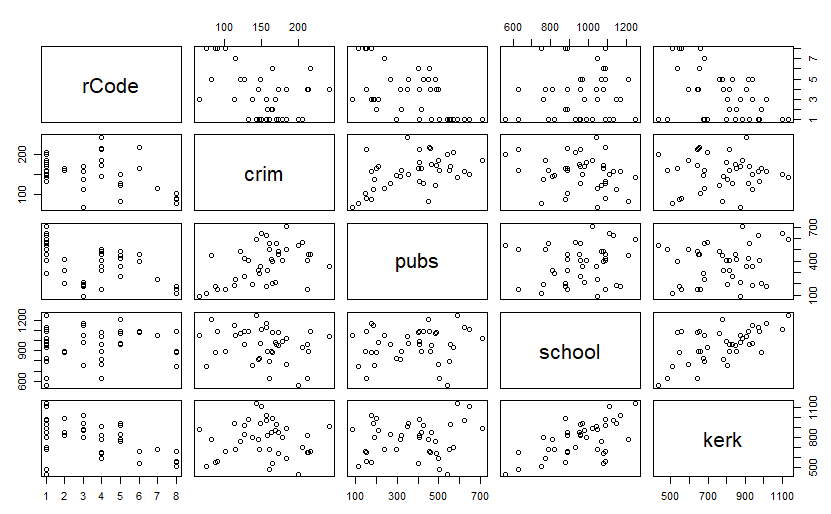
**1b)** Verwijder de variabelen Region en County. Deze nemen we niet mee met de analyse.

**Gevraagde antwoorden:**

Geef alle gebruikte R-commando’s.

> data$readcounty <- NULL

> data$region <- NULL

**Vraag 2 (crimepubs.csv)  
2a)** Plot alle data om alvast een idee te krijgen van eventuele lineaire verbanden. Bepaal ook de correlatie tussen de variabelen.  
**Gevraagde antwoorden:**Geef alle gebruikte R-commando’s. Geef ook de plot en de uitvoer voor de correlatie.

> plot(data)

**2b)** Bespreek de resultaten. Welke verbanden zou je op basis hiervan kunnen vermoeden?  
**Gevraagde antwoorden:**Verband tussen school en kerk, en crim en pubs

**Vraag 3 (crimepubs.csv)**  
Bouw een lineair-regressie model voor deze data. De afhankelijke variabele is crim, in eerste instantie gaan we uit van alle andere variabelen als onafhankelijke variabelen.

**3a)** Bouw een lineair regressie model zoals hierboven beschreven. Geef de summary van dit model.  
**Gevraagde antwoorden:**Geef alle gebruikte R-commando’s. Geef ook de uitvoer.

> model=lm(crim~rCode+pubs+school+kerk, data=data)

> summary(model)

Call:

lm(formula = crim ~ rCode + pubs + school + kerk, data = data)

Residuals:

Min 1Q Median 3Q Max

-53.353 -20.504 -7.884 18.615 96.572

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 186.70797 43.24103 4.318 0.000123 \*\*\*

rCode -2.42587 4.15854 -0.583 0.563401

pubs 0.10539 0.04684 2.250 0.030827 \*

school -0.08153 0.05575 -1.462 0.152583

kerk 0.01684 0.05678 0.297 0.768512

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.91 on 35 degrees of freedom

Multiple R-squared: 0.3252, Adjusted R-squared: 0.2481

F-statistic: 4.217 on 4 and 35 DF, p-value: 0.006854

**3b)** Laat nu **één voor een** de niet-significante variabelen weg totdat je een model hebt met daarin alleen nog significante variabelen. Geef de summary van dit model.

**Gevraagde antwoorden:**Geef alle gebruikte R-commando’s. Geef ook de uitvoer. Laat zien in welke volgorde je de niet-significante variabelen weglaat.

**3c)** Bespreek de resultaten. Wat is je conclusie?

**Gevraagde antwoorden:**

Beantwoord de vraag.

> model=lm(crim~rCode+pubs+school, data=data)

> summary(model)

Call:

lm(formula = crim ~ rCode + pubs + school, data = data)

Residuals:

Min 1Q Median 3Q Max

-54.832 -19.047 -8.006 18.141 98.057

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 192.84555 37.48479 5.145 9.65e-06 \*\*\*

rCode -3.26229 3.01747 -1.081 0.2868

pubs 0.09991 0.04248 2.352 0.0243 \*

school -0.06906 0.03616 -1.910 0.0641 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.46 on 36 degrees of freedom

Multiple R-squared: 0.3235, Adjusted R-squared: 0.2671

F-statistic: 5.739 on 3 and 36 DF, p-value: 0.002575

> model=lm(crim~rCode+pubs, data=data)

> summary(model)

Call:

lm(formula = crim ~ rCode + pubs, data = data)

Residuals:

Min 1Q Median 3Q Max

-69.234 -25.145 -4.899 19.398 92.540

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 137.13999 24.37533 5.626 2.01e-06 \*\*\*

rCode -4.36066 3.06633 -1.422 0.1634

pubs 0.08218 0.04292 1.915 0.0633 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.7 on 37 degrees of freedom

Multiple R-squared: 0.255, Adjusted R-squared: 0.2147

F-statistic: 6.331 on 2 and 37 DF, p-value: 0.004318

> model=lm(crim~+pubs+school, data=data)

> summary(model)

Call:

lm(formula = crim ~ +pubs + school, data = data)

Residuals:

Min 1Q Median 3Q Max

-59.97 -22.93 -5.92 18.77 97.53

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 178.81261 35.24595 5.073 1.12e-05 \*\*\*

pubs 0.12635 0.03482 3.629 0.000854 \*\*\*

school -0.07651 0.03557 -2.151 0.038108 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35.54 on 37 degrees of freedom

Multiple R-squared: 0.3016, Adjusted R-squared: 0.2638

F-statistic: 7.987 on 2 and 37 DF, p-value: 0.001308

**Vraag 4 (crime2.csv)**Gebruik lineaire regressie voor een onderzoek naar het verschil tussen gewelddadige misdaad en misdaad in het algemeen. Gebruik hiervoor de tweede dataset (crime2.csv).

De hoofdvraag is : Zijn er verschillen tussen gewelddadige misdaad en misdaad in het algemeen?

Subvragen zijn wat de bepalende factoren zijn hierbij.

**4a)** Uit de beschrijving blijkt al dat variabelen X1 en X2 met elkaar samenhangen. Plot deze data en bepaal de correlatie tussen X1 en X2 om dit duidelijk te maken.

**Gevraagde antwoorden:**Geef de R-commando’s en de uitvoer, inclusief de plot.

> setwd("D:/GitHub/School/P2.3/Analytics/opgaven\_data")

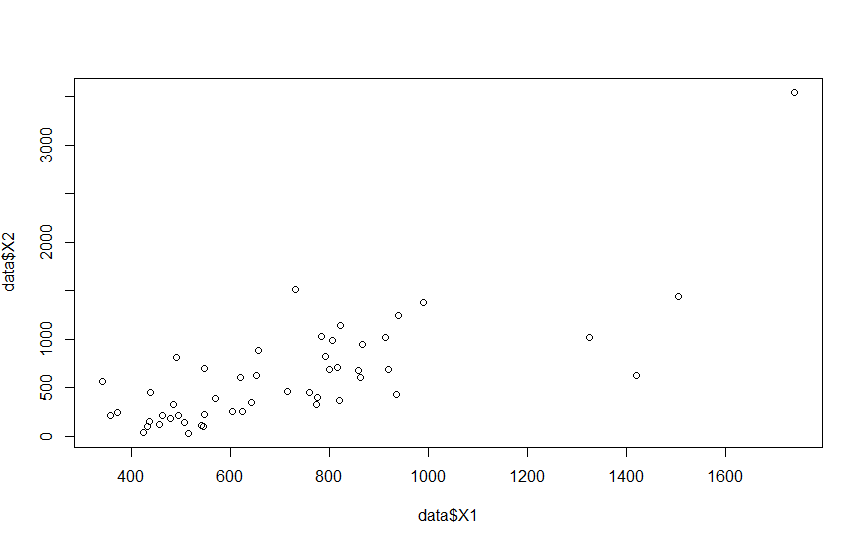
> data = read.csv("Crime2.csv")

> plot(data)

> plot(data$X1,data$X2)

> cor(data$X1, data$X2)

[1] 0.7565051



**4b)** Welke conclusie kan je uit de correlatie en de plot trekken? Waarom is een model met zowel X1 als X2 niet zinvol?

**Gevraagde antwoorden:**

Beantwoord de vragen.

Er is een duidelijke correlatie X1 en 2. De correlatie is hoog genoeg dat 1 van de 2 triviaal word, welke maakt niet uit.

Voor het vervolg gaan we een model maken van X1 als functie van X3 t/m X7, en ook een model van X2 als functie van X3 t/m X7.

**4c)** Maak een lineair regressiemodel van X1 als functie van X3 t/m X7. Verwijder daarna **steeds 1 variabele** totdat alleen significante variabelen overblijven. Geef de summary van het uiteindelijke model.

**Gevraagde antwoorden:**Geef alle gebruikte R-commando’s. Geef ook de uitvoer. Laat zien in welke volgorde je de niet-significante variabelen weglaat.

> model1=lm(X1~X3+X4+X5+X6+X7, data=data)

> summary(model1)

Call:

lm(formula = X1 ~ X3 + X4 + X5 + X6 + X7, data = data)

Residuals:

Min 1Q Median 3Q Max

-346.60 -148.46 -62.43 111.75 788.21

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 489.649 472.366 1.037 0.305592

X3 10.981 3.078 3.568 0.000884 \*\*\*

X4 -6.088 6.544 -0.930 0.357219

X5 5.480 10.053 0.545 0.588428

X6 0.377 4.417 0.085 0.932367

X7 5.500 13.754 0.400 0.691150

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 253.2 on 44 degrees of freedom

Multiple R-squared: 0.3336, Adjusted R-squared: 0.2579

F-statistic: 4.405 on 5 and 44 DF, p-value: 0.002444

>model1=lm(X1~X3+X4+X5+X7, data=data)

> summary(model1)

Call:

lm(formula = X1 ~ X3 + X4 + X5 + X7, data = data)

Residuals:

Min 1Q Median 3Q Max

-349.43 -144.96 -61.09 111.76 789.96

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 518.506 326.229 1.589 0.118974

X3 10.925 2.973 3.674 0.000633 \*\*\*

X4 -6.415 5.253 -1.221 0.228427

X5 5.002 8.256 0.606 0.547652

X7 6.302 9.942 0.634 0.529395

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 250.4 on 45 degrees of freedom

Multiple R-squared: 0.3335, Adjusted R-squared: 0.2742

F-statistic: 5.629 on 4 and 45 DF, p-value: 0.000923

> model1=lm(X1~X3+X4+X7, data=data)

> summary(model1)

Call:

lm(formula = X1 ~ X3 + X4 + X7, data = data)

Residuals:

Min 1Q Median 3Q Max

-360.97 -147.07 -55.85 100.11 785.10

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 654.564 234.996 2.785 0.00774 \*\*

X3 11.786 2.594 4.544 3.99e-05 \*\*\*

X4 -7.549 4.875 -1.549 0.12832

X7 4.504 9.424 0.478 0.63493

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 248.7 on 46 degrees of freedom

Multiple R-squared: 0.3281, Adjusted R-squared: 0.2842

F-statistic: 7.486 on 3 and 46 DF, p-value: 0.0003503

> model1=lm(X1~X3+X4, data=data)

> summary(model1)

Call:

lm(formula = X1 ~ X3 + X4, data = data)

Residuals:

Min 1Q Median 3Q Max

-356.90 -162.92 -60.86 100.69 784.30

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 621.426 222.685 2.791 0.00758 \*\*

X3 11.858 2.568 4.618 3.02e-05 \*\*\*

X4 -5.973 3.561 -1.677 0.10013

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 246.6 on 47 degrees of freedom

Multiple R-squared: 0.3247, Adjusted R-squared: 0.296

F-statistic: 11.3 on 2 and 47 DF, p-value: 9.838e-05

**4d)** Maak een lineair regressiemodel van X2 als functie van X3 t/m X7. Verwijder daarna **steeds 1 variabele** totdat alleen significante variabelen overblijven. Geef de summary van het uiteindelijke model.

**Gevraagde antwoorden:**Geef alle gebruikte R-commando’s. Geef ook de uitvoer. Laat zien in welke volgorde je de niet-significante variabelen weglaat.

model2=lm(X2~X3+X4+X5+X6+X7, data=data)

> summary(model2)

Call:

lm(formula = X2 ~ X3 + X4 + X5 + X6 + X7, data = data)

Residuals:

Min 1Q Median 3Q Max

-906.69 -299.90 -87.02 179.90 1929.51

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1171.268 920.598 1.272 0.20995

X3 21.010 5.998 3.503 0.00107 \*\*

X4 -23.911 12.753 -1.875 0.06745 .

X5 -7.097 19.593 -0.362 0.71893

X6 -6.565 8.609 -0.763 0.44980

X7 26.273 26.805 0.980 0.33236

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 493.5 on 44 degrees of freedom

Multiple R-squared: 0.3356, Adjusted R-squared: 0.2601

F-statistic: 4.446 on 5 and 44 DF, p-value: 0.002303

> model2=lm(X2~X3+X4+X5+X7, data=data)

> summary(model2)

Call:

lm(formula = X2 ~ X3 + X4 + X5 + X7, data = data)

Residuals:

Min 1Q Median 3Q Max

-809.02 -303.88 -83.46 200.59 1903.89

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 668.826 639.924 1.045 0.301529

X3 21.987 5.833 3.770 0.000474 \*\*\*

X4 -18.233 10.305 -1.769 0.083621 .

X5 1.226 16.196 0.076 0.939987

X7 12.325 19.502 0.632 0.530605

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 491.2 on 45 degrees of freedom

Multiple R-squared: 0.3269, Adjusted R-squared: 0.267

F-statistic: 5.463 on 4 and 45 DF, p-value: 0.001133

> model2=lm(X2~X3+X4+X7, data=data)

> summary(model2)

Call:

lm(formula = X2 ~ X3 + X4 + X7, data = data)

Residuals:

Min 1Q Median 3Q Max

-814.88 -305.33 -84.35 199.02 1903.21

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 702.176 459.124 1.529 0.1330

X3 22.198 5.068 4.380 6.79e-05 \*\*\*

X4 -18.511 9.524 -1.944 0.0581 .

X7 11.884 18.411 0.645 0.5218

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 485.9 on 46 degrees of freedom

Multiple R-squared: 0.3268, Adjusted R-squared: 0.2829

F-statistic: 7.443 on 3 and 46 DF, p-value: 0.0003653

> model2=lm(X2~X3+X4, data=data)

> summary(model2)

Call:

lm(formula = X2 ~ X3 + X4, data = data)

Residuals:

Min 1Q Median 3Q Max

-754.38 -352.37 -92.48 198.00 1894.68

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 614.744 435.957 1.410 0.1651

X3 22.389 5.027 4.453 5.19e-05 \*\*\*

X4 -14.354 6.972 -2.059 0.0451 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 482.8 on 47 degrees of freedom

Multiple R-squared: 0.3207, Adjusted R-squared: 0.2918

F-statistic: 11.09 on 2 and 47 DF, p-value: 0.0001132

**4e)** Bespreek de resultaten. Wat is volgens deze modellen het belangrijkste verschil tussen X1 (algehele misdaad) en X2 (gewelddadige misdaad)?

**Gevraagde antwoorden:**Beantwoord de vraag.

Geweldadige misdaag hangt van een extra variable af (X4)