# 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.

library(readr)

crimepubs <- read\_csv("analytics/opgaven\_data/crimepubs.csv")

View(crimepubs)

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

**Gevraagde antwoorden:**

Geef alle gebruikte R-commando’s.

crimepubs$county <- NULL

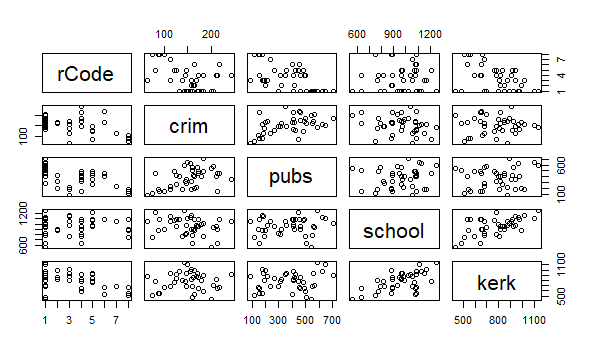
crimepubs$region <- NULL

**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(crimepubs)

**2b)** Bespreek de resultaten. Welke verbanden zou je op basis hiervan kunnen vermoeden?  
**Gevraagde antwoorden:**Beantwoord de vraag.

Er is een lineare correlatie te zien tussen crim en pubs en ook tussen school en kerk. De andere stippen op de plot zijn verspreid en zullen een lage correlatie hebben.



**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.

Punt onafhankelijk

**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.

Plot(crimepubs)

model=lm(crim~pubs,data=crimepubs)

abline(model,col="red")

summary(model)

Call:

lm(formula = crim ~ pubs, data = crimepubs)

Residuals:

Min 1Q Median 3Q Max

-78.517 -25.704 -1.653 22.592 90.988

Coefficients:

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

(Intercept) 109.3400 14.7553 7.410 6.9e-09 \*\*\*

pubs 0.1162 0.0361 3.219 0.00263 \*\*

---

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

Residual standard error: 37.19 on 38 degrees of freedom

Multiple R-squared: 0.2142, Adjusted R-squared: 0.1936

F-statistic: 10.36 on 1 and 38 DF, p-value: 0.002635

model=lm(crim~school,data=crimepubs)

abline(model,col="red")

summary(model)

Call:

lm(formula = crim ~ school, data = crimepubs)

Residuals:

Min 1Q Median 3Q Max

-89.160 -24.648 5.593 20.639 92.954

Coefficients:

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

(Intercept) 209.42199 39.32237 5.326 4.78e-06 \*\*\*

school -0.05902 0.04050 -1.457 0.153

---

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

Residual standard error: 40.83 on 38 degrees of freedom

Multiple R-squared: 0.05292, Adjusted R-squared: 0.028

F-statistic: 2.123 on 1 and 38 DF, p-value: 0.1533

model=lm(crim~kerk,data=crimepubs)

abline(model,col="red")

summary(model)

Call:

lm(formula = crim ~ kerk, data = crimepubs)

Residuals:

Min 1Q Median 3Q Max

-86.982 -25.964 4.483 21.303 87.989

Coefficients:

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

(Intercept) 1.522e+02 3.111e+01 4.894 1.85e-05 \*\*\*

kerk 8.556e-04 3.895e-02 0.022 0.983

---

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

Residual standard error: 41.96 on 38 degrees of freedom

Multiple R-squared: 1.269e-05, Adjusted R-squared: -0.0263

F-statistic: 0.0004824 on 1 and 38 DF, p-value: 0.9826

**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.

crimepubs$school <- NULL

crimepubs$kerk <- NULL

model=lm(crim~pubs,data=crimepubs)

abline(model,col="red")

summary(model)

Call:

lm(formula = crim ~ pubs, data = crimepubs)

Residuals:

Min 1Q Median 3Q Max

-78.517 -25.704 -1.653 22.592 90.988

Coefficients:

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

(Intercept) 109.3400 14.7553 7.410 6.9e-09 \*\*\*

pubs 0.1162 0.0361 3.219 0.00263 \*\*

---

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

Residual standard error: 37.19 on 38 degrees of freedom

Multiple R-squared: 0.2142, Adjusted R-squared: 0.1936

F-statistic: 10.36 on 1 and 38 DF, p-value: 0.002635

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

**Gevraagde antwoorden:**

Beantwoord de vraag.

Alleen crim met pubs lijkt significant te zijn, crim met school of kerk niet. Waarvan school het minst significant lijkt te zijn, dus deze wordt eerst weggelaten.

**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.

> library(readr)

> crime <- read\_csv("analytics/opgaven\_data/Crime2.csv")

> View(crime)

> model=lm(X1~X2,data=crime)

> plot(model)

> abline(model,col="red")

> summary(model)

Call:

lm(formula = X1 ~ X2, data = crime)

Residuals:

Min 1Q Median 3Q Max

-357.12 -98.58 -22.17 56.87 695.30

Coefficients:

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

(Intercept) 479.14487 40.52693 11.823 7.99e-16 \*\*\*

X2 0.38757 0.04836 8.014 2.10e-10 \*\*\*

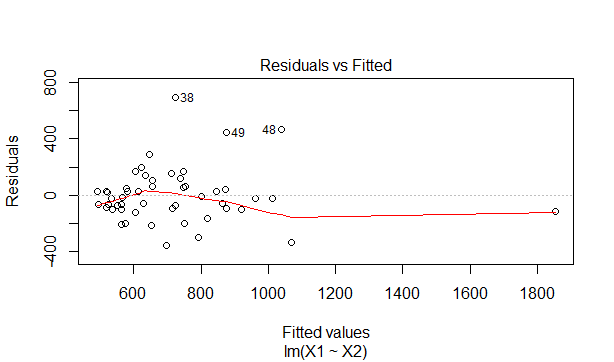
---

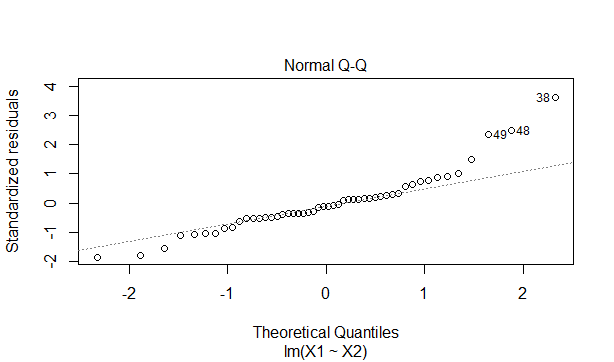
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

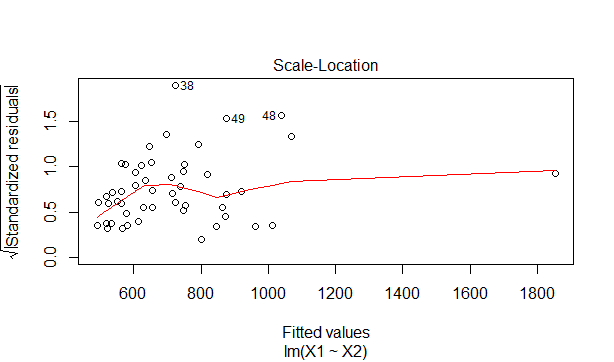
Residual standard error: 194.2 on 48 degrees of freedom

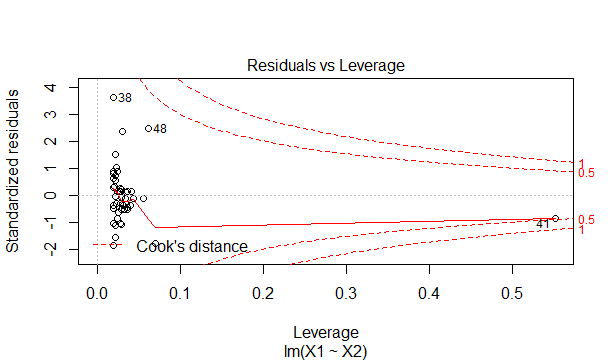
Multiple R-squared: 0.5723, Adjusted R-squared: 0.5634

F-statistic: 64.23 on 1 and 48 DF, p-value: 2.096e-10







****

**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.

De correlatie tussen X1 en X2 is significant zoals ook te zien is bij normal QQ plot. Een model met beide is niet zinvol omdat je ook wilt weten of X3 t/m X7 significante factor is.

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.

> model=lm(X1~X3,data=crime)

> summary(model)

Call:

lm(formula = X1 ~ X3, data = crime)

Residuals:

Min 1Q Median 3Q Max

-329.66 -175.91 -66.84 137.48 794.66

Coefficients:

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

(Intercept) 289.749 104.304 2.778 0.00778 \*\*

X3 11.340 2.597 4.367 6.7e-05 \*\*\*

---

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

Residual standard error: 251.2 on 48 degrees of freedom

Multiple R-squared: 0.2843, Adjusted R-squared: 0.2694

F-statistic: 19.07 on 1 and 48 DF, p-value: **6.699e-05**

> model=lm(X1~X4,data=crime)

> summary(model)

Call:

lm(formula = X1 ~ X4, data = crime)

Residuals:

Min 1Q Median 3Q Max

-380.14 -190.45 -36.16 105.31 1034.83

Coefficients:

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

(Intercept) 952.899 251.496 3.789 0.000422 \*\*\*

X4 -3.996 4.218 -0.947 0.348271

---

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

Residual standard error: 294.2 on 48 degrees of freedom

Multiple R-squared: 0.01835, Adjusted R-squared: -0.002102

F-statistic: 0.8972 on 1 and 48 DF, p-value: **0.3483**

> model=lm(X1~X5,data=crime)

> summary(model)

Call:

lm(formula = X1 ~ X5, data = crime)

Residuals:

Min 1Q Median 3Q Max

-338.9 -169.9 -61.9 108.3 918.2

Coefficients:

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

(Intercept) 475.598 110.097 4.320 7.8e-05 \*\*\*

X5 15.738 6.667 2.361 0.0224 \*

---

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

Residual standard error: 281.1 on 48 degrees of freedom

Multiple R-squared: 0.104, Adjusted R-squared: 0.08535

F-statistic: 5.573 on 1 and 48 DF, p-value: **0.02235**

> model=lm(X1~X6,data=crime)

> summary(model)

Call:

lm(formula = X1 ~ X6, data = crime)

Residuals:

Min 1Q Median 3Q Max

-394.01 -198.24 -77.59 124.92 980.63

Coefficients:

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

(Intercept) 822.007 93.968 8.748 1.67e-11 \*\*\*

X6 -3.480 2.822 -1.233 0.224

---

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

Residual standard error: 292.4 on 48 degrees of freedom

Multiple R-squared: 0.0307, Adjusted R-squared: 0.01051

F-statistic: 1.52 on 1 and 48 DF, p-value: **0.2236**

> model=lm(X1~X7,data=crime)

> summary(model)

Call:

lm(formula = X1 ~ X7, data = crime)

Residuals:

Min 1Q Median 3Q Max

-369.2 -218.8 -60.2 125.8 1023.8

Coefficients:

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

(Intercept) 738.661 121.154 6.097 1.78e-07 \*\*\*

X7 -1.498 8.223 -0.182 0.856

---

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

Residual standard error: 296.9 on 48 degrees of freedom

Multiple R-squared: 0.0006908, Adjusted R-squared: -0.02013

F-statistic: 0.03318 on 1 and 48 DF, p-value: **0.8562**

crime$X7 <- NULL

crime$X4 <- NULL

crime$X6 <- NULL

**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.

> model=lm(X2~X3,data=crime)

> summary(model)

Call:

lm(formula = X2 ~ X3, data = crime)

Residuals:

Min 1Q Median 3Q Max

-907.45 -333.10 -58.11 198.93 1908.80

Coefficients:

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

(Intercept) -182.245 207.098 -0.880 0.383246

X3 21.145 5.157 4.101 0.000158 \*\*\*

---

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

Residual standard error: 498.9 on 48 degrees of freedom

Multiple R-squared: 0.2594, Adjusted R-squared: 0.244

F-statistic: 16.81 on 1 and 48 DF, p-value: **0.0001583**

> model=lm(X2~X4,data=crime)

> summary(model)

Call:

lm(formula = X2 ~ X4, data = crime)

Residuals:

Min 1Q Median 3Q Max

-568.0 -313.2 -144.5 220.0 2962.8

Coefficients:

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

(Intercept) 1240.591 486.962 2.548 0.0141 \*

X4 -10.619 8.168 -1.300 0.1997

---

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

Residual standard error: 569.7 on 48 degrees of freedom

Multiple R-squared: 0.03402, Adjusted R-squared: 0.0139

F-statistic: 1.69 on 1 and 48 DF, p-value: **0.1997**

> model=lm(X2~X5,data=crime)

> summary(model)

Call:

lm(formula = X2 ~ X5, data = crime)

Residuals:

Min 1Q Median 3Q Max

-680.8 -325.3 -194.0 239.9 2745.9

Coefficients:

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

(Intercept) 189.30 217.20 0.872 0.3878

X5 27.72 13.15 2.108 0.0403 \*

---

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

Residual standard error: 554.6 on 48 degrees of freedom

Multiple R-squared: 0.0847, Adjusted R-squared: 0.06563

F-statistic: 4.442 on 1 and 48 DF, p-value: **0.04032**

> model=lm(X2~X6,data=crime)

> summary(model)

Call:

lm(formula = X2 ~ X6, data = crime)

Residuals:

Min 1Q Median 3Q Max

-616.0 -339.2 -123.7 225.7 2837.0

Coefficients:

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

(Intercept) 846.907 182.569 4.639 2.73e-05 \*\*\*

X6 -7.717 5.483 -1.407 0.166

---

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

Residual standard error: 568.1 on 48 degrees of freedom

Multiple R-squared: 0.03963, Adjusted R-squared: 0.01962

F-statistic: 1.981 on 1 and 48 DF, p-value: **0.1658**

> model=lm(X2~X7,data=crime)

> summary(model)

Call:

lm(formula = X2 ~ X7, data = crime)

Residuals:

Min 1Q Median 3Q Max

-601.5 -380.6 -151.6 202.0 2934.8

Coefficients:

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

(Intercept) 686.239 236.317 2.904 0.00556 \*\*

X7 -5.069 16.040 -0.316 0.75334

---

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

Residual standard error: 579.1 on 48 degrees of freedom

Multiple R-squared: 0.002077, Adjusted R-squared: -0.01871

F-statistic: 0.09989 on 1 and 48 DF, p-value: **0.7533**

crime$X7 <- NULL

crime$X4 <- NULL

crime$X6 <- NULL

**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.

Zowel factor X7, X4 en X6 zijn niet significant bij X1 en X2. X3 en X5 zijn wel significant.