Computerzitting 3:
Testen op onafhankelijkheid in R
2022-2023

Om na te gaan of 2 variabelen onafhankelijk zijn, moet er eerst een onderscheid gemaakt worden tussen continue en discrete variabelen. We gebruiken immers een andere methode om te testen op onafhankelijkheid voor de discrete ten opzichte van de continue variabelen. De χ^2-test is geschikt voor de discrete variabelen, waar lineaire regressie en de test op correlatie gebruikt kunnen worden voor continue variabelen.

1 Discrete variabelen: χ^2-test

Voor de jaarlijkse sportdag kunnen de leerlingen kiezen uit 3 sporten: atletiek, paardrijden of ijsschaatsen. De sportleerkrachten zijn benieuwd of er een verband is tussen de gekozen sport en het geslacht (jongen of meisje) van de leerling. De geobserveerde contingentietabel is

	atletiek	paardrijden	schaatsen	
jongen	$34 = f_{11}$	$33 = f_{12}$	$7 = f_{13}$	$74 = f_{x1}$
meisje	$22 = f_{21}$	$22 = f_{22}$	$4 = f_{23}$	$48 = f_{x2}$
	$56 = f_{y1}$	$55 = f_{y2}$	$11 = f_{y3}$	$122 = n$

Om een antwoord te geven op de vraag van de sportleerkrachten, voeren we een χ^2-toets uit voor de onafhankelijkheid. We vergelijken de geobserveerde aantallen met de aantallen die we zouden verwachten als er geen verband zou zijn tussen geslacht en gekozen sport. Dan zou

$$p_{ij} = p_i p_j \qquad i = 1, 2 \quad j = 1, 2, 3$$

met p_{ij} de kans om tot de i-de rij en de j-de kolom te behoren, p_i de kans om tot de i-de rij te behoren en p_j de kans om tot de j-de kolom te behoren.

De kansen benader je door de relatieve frequenties. Dus :

$$\frac{f_{ij}}{n} \approx \frac{f_{xi}}{n}\frac{f_{yj}}{n}$$
$$f_{ij} \approx \frac{f_{xi} f_{yj}}{n}$$

De verwachte aantallen zijn dan gelijk aan :

	atletiek	paardrijden	schaatsen	
jongen	33.967	33.361	6.672	$74 = f_{x1}$
meisje	22.033	21.639	4.328	$48 = f_{x2}$
	$56 = f_{y1}$	$55 = f_{y2}$	$11 = f_{y3}$	$122 = n$

We zien dat de verwachten aantallen niet zoveel afwijken van de geobeserveerde aantallen. In wat volgt gaan we op een formele wijze na of de afwijkingen al dan niet te groot zijn. Indien de afwijkingen te groot zijn, zal de veronderstelling dat er geen verband is tussen geslacht en sport niet kloppen.

Om afhankelijkheid na te gaan, wordt de volgende hypothese getest:

$$H_0 : \text{Het geslacht is onafhankelijk van de gekozen sport.}$$
$$\text{versus}$$
$$H_1 : \text{Er is een verband tussen het geslacht en de gekozen sport.}$$

Om deze hypothese te testen, berekenen we het χ^2-getal, met behulp van de volgende formule :

$$\sum_{\text{alle cellen}} \frac{(\text{geobserveerde waarde - verwachte waarde})^2}{\text{verwachte waarde}} = 0.0509.$$

Dit getal kunnen we ook met R berekenen. Ga hiervoor als volgt te werk:

1. Creëer de contingentietabel, i.e. een matrix met de ene variabele (vb. de verschillende sporten) op de rijen en de andere variabele (het geslacht) in de kolommen:

   ```
   NrJongens = c(34,33,7)
   NrMeisjes = c(22,22,4)
   Ctable = data.frame(NrJongens,NrMeisjes)
   ```

2. Voer de χ^2 test uit:

   ```
   ChiSq = chisq.test(Ctable)
   ChiSq
   ChiSq$observed
   ChiSq$expected
   ```

3. Herhaal deze procedure, maar nu met een contingentietabel waarbij geslacht de rijen zijn en sport de kolommen. Vanzelfsprekend moet je exact dezelfde resultaten bekomen.

In de uitvoer kan je de waarde van de teststatistiek 0.0509 en de p-waarde 0.9749 terugvinden.

Merk op: er verschijnt een waarschuwing in het commando window dat de χ^2 benadering niet nauwkeurig kan zijn. Dit komt omdat 1 cel (Meisjes-Schaatsen) minder dan 5 observaties bevat.

Intuïtief voelen we aan dat hoe kleiner het χ^2-getal is (m.a.w. hoe dichter de geobserveerde aantallen bij de verwachte aantallen liggen), hoe meer waarschijnlijk H_0 zal zijn.

We kunnen ook formeel besluiten of we H_0 al dan niet verwerpen met behulp van de p-waarde op significantieniveau $\alpha = 0.05$.

Wat is de p-waarde in dit voorbeeld?

.....

Wat besluit je? Indien je besluit dat er afhankelijkheid is, leg uit!

.....

.....

Veronderstel nu dat we de gegevens niet in een tabel gegeven krijgen, maar in een dataset die bestaat uit 2 kolommen, *geslacht* en *sport*, en 122 rijen zoals in de dataset *sport.xls* die op Blackboard te vinden is. We beschikken dan nog steeds over 2 categorische variabelen, waarvan we de onafhankelijkheid kunnen testen door middel van de χ^2-test. Dit gebeurt in R als volgt:

```
gegevens=read.csv(file=file.choose(),header=TRUE,dec=",",sep=";")

Ctable = ftable(gegevens)

ChiSq = chisq.test(Ctable)
ChiSq
ChiSq$observed
ChiSq$expected
```

Uiteraard vinden we opnieuw dezelfde waarde voor de teststatistiek en de p-waarde, en kan hetzelfde besluit als hierboven getrokken worden.

Oefening 1 We hernemen de eerste dataset uit Computerzitting 0 : *resultaten.xls*.

- Kiezen vrouwelijke studenten een andere kleur voor hun auto dan mannelijke studenten?

  .....

  .....

- Verschilt de keuze van studierichting voor mannen en vrouwen? Zo ja, leg uit.

  .....

  .....

2 Continue variabelen

Voor continue variabelen is het moeilijk om onafhankelijkheid na te gaan. Wel kunnen we nagaan of ze gecorreleerd zijn of niet, zowel op een grafische manier als op een formele manier. Door een scatterplot te maken kunnen we zien of er een (al dan niet lineair) verband bestaat tussen de twee variabelen. Daarnaast is er ook een formele test die gebruik maakt van de correlatiecoëfficiënt tussen de twee variabelen.

2.1 Correlatie analyse

Gegeven twee metrische variabelen X en Y, willen we de volgende hypothese onderzoeken:

$$H_0: \text{ er is geen lineaire afhankelijkheid tussen } X \text{ en } Y$$
$$H_1: \text{ er is een mate van lineaire afhankelijkheid tussen } X \text{ en } Y.$$

Dit is equivalent met de hypothese

$$H_0 : \rho = 0 \text{ versus } H_1 : \rho \neq 0$$

met ρ de populatiecorrelatiecoëfficiënt.

Om deze hypothese te kunnen testen, moeten we onderscheid maken tussen twee gevallen:

1. X en Y zijn (bivariaat) normaal verdeeld
2. X en Y zijn niet bivariaat normaal verdeeld.

2.1.1 X en Y zijn bivariaat normaal verdeeld

De hypothese

$$H_0 : \rho = 0 \text{ versus } H_1 : \rho \neq 0$$

wordt getest met behulp van de teststatistiek

$$T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim_{H_0} t_{n-2}.$$

Om deze test te kunnen uitvoeren, moet eerst de **Pearson** correlatiecoëfficiënt r_n berekend worden

$$r_n = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

voor de gegeven dataset (x_i, y_i). Deze correlatiecoëfficiënt ligt steeds tussen -1 en 1. Dan wordt t, de waarde voor de teststatistiek, berekend:

$$t = \frac{r_n\sqrt{n-2}}{\sqrt{1-r_n^2}}$$

met de bijbehorende p-waarde. Opnieuw zal de p-waarde vergeleken worden met het significantieniveau $\alpha = 0.05$ om een besluit te trekken.

Merk op dat we deze test enkel kunnen uitvoeren indien X en Y bivariaat normaal verdeeld zijn. Een nodige voorwaarde is dat X en Y elk afzonderlijk normaal verdeeld zijn. Een scatterplot kan helpen om te zien of de bivariate normaliteit plausibel lijkt. De gegevens moeten dan een elliptische wolk vormen.

Oefening 2 We halen de dataset *roken.xls* van Blackboard en importeren deze in R. Deze bevat informatie over het aantal rokers en de mortaliteit voor een aantal beroepsgroepen. De dataset bevat 3 variabelen: *Occupational_group*, *rokers* en *mortaliteit*.

1. Maak eerst een scatterplot van de data.

```
gegevens=read.csv(file=file.choose(),header=TRUE,dec=",",sep=";")

Rokers= gegevens$rokers
Mortaliteit = gegevens$mortaliteit
RokersMort = data.frame(Rokers,Mortaliteit)
plot(RokersMort,type="p",xlab="Aantal rokers",
     ylab="Mortaliteit",main="Mortaliteit = f(Aantal rokers)")
```

   Er is één observatie die duidelijk afwijkt van de rest, zo'n observatie noemen we een uitschieter. Een uitschieter kan de resultaten van de analyse heel sterk beïnvloeden. Daarom maken we een nieuwe dataset aan, waarbij de uitschieter wordt weggelaten met behulp van de code:

   - Detecteer de outlier (i.e. ga na welke observatie uit de originele dataset de outlier is):
     - Typ het commando
       `index_outliers<-identify(RokersMort,n=1)`
     - R brengt nu de figuur naar de voorgrond: klik op de observatie waarvan je vermoedt dat het een outlier is. R zal vervolgens tonen wat de (rij-)index van deze observatie is.
   - Verwijder de outlier d.m.v. het commando

     `RokersMortWithout = RokersMort[-index_outliers,]`

   Tip: Indien je meerdere outliers wilt detecteren en/of verwijderen, kan je in het algemeen als volgt te werk gaan:

   - Detecteer $k \geq 1$ outliers, via

     `index_outliers<-identify(RokersMort,n=k)`

     en klik k observaties aan op de scatterplot.
   - Verwijder deze k outliers via

     `RokersMortWithout = RokersMort[-index_outliers,]`

- De afzonderlijke variabelen rokers en mortaliteit met hun respectievelijke observatie(s) verwijderd, kan je bekomen via

  ```
  RokersWithout = RokersMortWithout$Rokers
  MortaliteitWithout = RokersMortWithout$Mortaliteit
  ```

2. Maak nu een scatterplot van de nieuwe dataset. Denk je dat een lineair verband plausibel is?

3. Ga na dat de variabelen *rokers* en *mortaliteit* normaal verdeeld zijn (na het verwijderen van de uitschieter!).

   .....

   .....

4. Bereken de Pearson correlatiecoëfficiënt r_n, de teststatistiek t en de bijhorende p-waarde met behulp van deze code:

   ```
   cor.test(RokersWithout,MortaliteitWithout,method="pearson")
   ```

   .....

5. Wat besluit je?

   .....

6. Klopt dit met wat je verwachtte op basis van de scatterplot?

   .....

Later in deze computerzitting zullen we dit lineair verband expliciet bepalen met behulp van lineaire regressie.

Oefening 3 De dataset *azijnzuur.xls* bevat de dichtheid van azijnzuur-watermengsels en het gewichtspercentage azijnzuur in het mengsel. Men wil weten of er een verband is tussen het gewichtspercentage en de dichtheid.

1. Welke waarde bekom je voor r_n, voor de teststatistiek t en wat is de bijbehorende p-waarde?

   .....

   .....

2. Stel dat je veronderstelt dat (X, Y) bivariaat normaal verdeeld zijn, zonder dit expliciet na te gaan. Wat zou je hieruit dan besluiten?

   .....

3. Onderzoek de bivariate normaliteit van de gegevens. Wat is je besluit nu?

   .....

2.1.2 X en Y zijn niet bivariaat normaal verdeeld

Als X en Y niet uit een bivariate normale verdeling komen, kan de test met behulp van de Pearson correlatiecoëfficiënt niet gebruikt worden. Men kan wel altijd de hypothese

$$H_0: \text{ er is geen monotoon verband tussen } X \text{ en } Y$$
$$H_1: \text{ er is een mate van monotoon verband tussen } X \text{ en } Y$$

testen met behulp van de Spearman correlatiecoëfficiënt. I.p.v. met de geobserveerde waarden te werken, gaat men met rangnummers te werk om de Spearman correlatiecoëfficiënt te berekenen. De Spearman correlatiecoëfficiënt is dan niets anders dan de Pearson correlatiecoëfficiënt maar nu berekend op basis van die rangnummers. De Spearman correlatiecoëfficiënt zal ook een waarde aannemen tussen -1 en 1. Een negatieve waarde voor deze Spearman correlatiecoëfficiënt duidt erop dat de ene variabele de neiging heeft te stijgen als de andere afneemt, en een positieve Spearman correlatiecoëfficiënt geeft aan dat de ene variabele toeneemt als de andere variabele stijgt. Indien H_0 verworpen wordt, en er dus een mate van monotoon verband is, kan men met behulp van lineaire regressie onderzoeken of het zinvol is om te stellen dat dit verband lineair is. Lineaire regressie veronderstelt immers geen normaliteit van X en Y, enkel van de foutentermen.

Oefening 4 Open de dataset *vis.xls*. Deze bevat 2 variabelen, Prijs70 en Prijs80.

1. Onderzoek de normaliteit van de variabelen *Prijs70* en *Prijs80*.

   .....

   .....

2. Welke waarde bekom je voor r_s en wat is de bijbehorende p-waarde?

   .....

3. Wat kan je besluiten?

   .....

4. Komt dit overeen met wat je ziet op de scatterplot?

   .....

Lineaire regressie zal ons helpen om de aard van dit verband te beschrijven.

2.2 Lineaire regressie

Stel dat we voor n onderzoekseenheden twee variabelen $\{x_1, \ldots, x_n \in \mathbb{R}\}$ en $\{y_1, \ldots, y_n \in \mathbb{R}\}$ gegeven hebben. Met lineaire regressie onderzoeken we of er een lineair verband is tussen de twee variabelen X en Y. We onderzoeken daartoe of volgend model gerechtvaardigd is:

$$y_i = a + b\ x_i + \epsilon_i$$

met ϵ_i onafhankelijk en verdeeld volgens $N(0, \sigma^2)$. De getallen $\epsilon_i = y_i - (a + b\ x_i)$ noemen we de foutentermen. Schattingen $\hat{a}$ en $\hat{b}$ van a en b vinden we met de kleinstekwadratenmethode: we zoeken a en b zodanig dat

$$\sum_{i=1}^{n} (y_i - (a + b\ x_i))^2$$

zo klein mogelijk is.
Met behulp van de residuen $r_i = y_i - (\hat{a} + \hat{b}\ x_i) = y_i - \hat{y}_i$ gaan we na of de foutentermen inderdaad onafhankelijk zijn en normaal verdeeld met zelfde variantie. Een schatting s^2 voor deze variantie is

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} r_i^2.$$

Als we besluiten dat onze aannames gerechtvaardigd zijn, kunnen we het verband tussen X en Y numeriek onderzoeken.
We kunnen daartoe eerst testen of b gelijk is aan nul. Als $b = 0$, is immers

$$y_i = a + \epsilon_i.$$

Bijgevolg is Y dan niet lineair afhankelijk van X. Dus een belangrijke hypothesetest is de test

$$H_0 : b = 0 \quad \text{versus} \quad H_1 : b \neq 0.$$

De waarde t voor teststatistiek wordt in dit geval gegeven door

$$t = \frac{\hat{b}}{\text{s.e.}(\hat{b})}$$

met $\text{s.e.}(\hat{b}) = s \dfrac{1}{\sqrt{(n-1)s_x^2}} = \dfrac{s}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$.
Onder H_0 komt t uit een student t-verdeling met $n - 2$ vrijheidsgraden.

Tenslotte bekijken we ook het getal R^2, een getal dat informatie geeft over de varianties in het model.
We vertrekken vanuit de volgende gelijkheid:

$$\sum_{i=1}^{n} r_i^2 + \sum_{i=1}^{n} (\overline{y} - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2.$$

of

$$SSE + SSR = SST.$$

In woorden betekent dit dat de onverklaarde variantie (SSE) plus de verklaarde variantie (SSR) gelijk is aan de totale variantie (SST). Definieer nu R^2 door

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n} (\overline{y} - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}.$$

We zien dat R^2 gelijk is aan de verklaarde variantie gedeeld door de totale variantie. Dus als R^2 dicht bij 1 ligt, is bijna alle variantie verklaard door het model. Dat betekent dat we een goed model hebben. Hoe verder R^2 van 1 ligt, hoe minder van de totale variantie verklaard wordt en hoe slechter het model. Toch is deze R^2 geen absolute maat, voorzichtigheid is geboden!

2.2.1 Voorbeeld

We keren terug naar de dataset *roken.xls* uit oefening 4.

Stap 1: onderzoeken of er een lineair of monotoon verband aanwezig is met behulp van r_n of r_s en een scatterplot

In oefening 2 hebben we op basis van de scatterplot en de Pearson correlatiecoëfficiënt besloten dat er een lineair verband is tussen de variabelen rokers en mortaliteit (na het verwijderen van de outlier).

Stap 2: opstellen van het regressiemodel

We voeren nu lineaire regressie uit op de nieuwe dataset om dit verband expliciet te bepalen. Het basisprogramma voor de regressie gaat als volgt:

```
Results = lm(MortaliteitWithout ~ RokersWithout)
summary(Results)
```

Dit programma schat de lineaire relatie $f(x) = a + bx$ tussen beide variabelen, als volgt: Mortaliteit = f(Rokers) $= a + b$Rokers. Als uitvoer van deze code vinden we

```
Call: lm(formula = MortaliteitWithout ~ RokersWithout)

Residuals:
Min      1Q       Median   3Q      Max
-30.369  -18.685  3.116    13.988  31.450

Coefficients:
            Estimate Std.Error t value Pr(>|t|)
(Intercept)  -2.4983  23.4971  -0.106  0.916
RokersWithout 1.0866   0.2252   4.825  8.05e-05  ***
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 18.97 on 22 degrees of freedom
Multiple R-squared: 0.5142,    Adjusted R-squared: 0.4921
F-statistic: 23.28 on 1 and 22 DF,  p-value: 8.048e-05
```

- Onder Coefficients vinden we

– *Estimate*: schattingen $\hat{a}$ (*intercept*) en $\hat{b}$ (*richtingscoëfficient*) van de parameters a en b met behulp van de kleinstekwadratenmethode. We krijgen $\hat{a} = -2.49825$ en $\hat{b} = 1.08662$. De regressierechte wordt dus

$$\text{Mortaliteit} = -2.4983 + 1.0866 * \text{Rokers} \ .$$

– t-value: geeft voor elke regressiecoëfficient $\beta_i, i \in \{0, 1\}$ met $\beta_0 = a$ en $\beta_1 = b$, de waarde van de teststatistiek m.b.t. de hypothesetest (t-test)

$$H_0 : \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0.$$

– $\Pr(> |t|)$: de p-waarde van de bovenvermelde hypothesetoets.

Het is vooral van belang om te kijken naar de hypothesetest $H_0 : b = 0$ versus $H_1 : b \neq 0$. Dit vertelt ons immers of ons model zinvol is (H_1) of niet (H_0). Immers, indien we H_0 niet kunnen verwerpen, dan is Mortaliteit=f(Rokers)=a en wordt de mortaliteit niet beïnvloed door het rookgedrag. In ons geval vinden we een waarde voor de teststatistiek $t = 4.825$, met bijhorende p-waarde gelijk aan $8.05e - 05 < 0.05$, zodat we de nulhypothese $H_0 : b = 0$ verwerpen op een significantieniveau van $\alpha = 0.05$. Het model is dus zinvol, en we mogen besluiten dat er een lineair verband bestaat tussen de variabelen mortaliteit en rokers.

- Residual standard error geeft een schatting s voor σ, i.e. de standaarddeviatie van de residuen. We vinden $s = 18.97$.

- R-squared is een maat voor de sterkte van de afhankelijkheid tussen mortaliteit en rokers. Hier vinden we $R^2 = 0.5142$. Er is dus een lineair verband maar met vrij veel ruis, waarbij 51% van de variabiliteit op de Mortaliteit verklaard wordt door het aantal rokers.

  Merk op dat de berekening van de correlatiecoëfficiënt en de enkelvoudige lineaire regressie niet los staan van elkaar. Immers, bij enkelvoudige lineaire regressie geldt dat $r_n = \sqrt{R^2}$. (In tegenstelling tot de gewone R^2 houdt de Adjusted R-squared R^2_{adj} rekening met het aantal verklarende variabelen in het model. Er geldt steeds: $R^2 \leq R^2_{adj}$.)

- F-statistic: een F-test om na te gaan of het model zinvol is. Deze test levert dezelfde p-waarde als bij t-test in het geval van enkelvoudige lineaire regressie. In het meervoudig geval niet noodzakelijk. Dan is dit de teststastiek die nagaat dat de vector met rico's (dus zonder intercept) gelijk is aan de nulvector.

Stap 3: nagaan van de modelonderstellingen

1. Het resultaat van deze testen is enkel relevant als aan de klassieke voorwaarden van lineaire regressie voldaan zijn. Om die na te gaan is een studie van de residuen vereist

   $$r_i = y_i - (\hat{a} + \hat{b}\ x_i) = y_i - \hat{y}_i,$$

   die een benadering zijn van de foutentermen ϵ_i. Volgens de klassieke voorwaarden zijn deze foutentermen onderling onafhankelijk en normaal verdeeld met dezelfde variantie. Hiertoe bekijken we de Normale QQ-plot van de residuen:

```
RawRes = residuals(Results)
qqnorm(RawRes,xlab="Standard Normal Quantiles",
     ylab="Raw Residuals",main="Raw Residual-Normal QQ-plot")
```

Wat besluit je over de residuen?

.....

Merk op dat we hier geen Shapiro-Wilk test uitvoeren omdat we niet geïnteresseerd zijn in de normaliteit van de residuen, maar wel in de normaliteit van de foutentermen ϵ_i.

2. Verder kunnen nog een aantal diagnostische plots bekeken worden.

```
plot(RokersWithout,MortaliteitWithout,xlab="Rokers",
   ylab="Mortaliteit",main="Mortaliteit = f(Rokers)")
abline(Results,col="red",lty=1)

StandRes = rstandard(Results)
plot(RokersWithout,StandRes,xlab="Rokers",
   ylab="Standardized Residuals",main="Standardized Residuals")
abline(h=0,col="blue",lty=1)
abline(h=-2.5,col="red",lty=1)
abline(h=2.5,col="red",lty=1)

predictedMort = predict.lm(Results)
plot(MortaliteitWithout,predictedMort,xlab="Geobserveerde Mortaliteit",
   ylab="Voorspelde Mortaliteit",
   main="Voorspelde Mortaliteit = f(Geobserveerde Mortaliteit)")
abline(a=0,b=1,col="red",lty=1)
```

Laten we deze figuren eens in wat meer detail bekijken.

- De eerste plot is de gewone scatterplot van beide variabelen die we ondertussen goed kennen, met daarbij ook de regressierechte.
- In de tweede grafiek zijn de gestandardizeerde residuen uitgezet ten opzichte van de verklarende variabele rokers. Opdat de klassieke voorwaarden voor regressie voldaan zouden zijn, verwacht je dat de puntenwolk symmetrisch gelegen is rond 0 en dat er geen patroon te herkennen valt in deze plot (i.e. constante variantie en onafhankelijkheid). In dit voorbeeld liggen de punten inderdaad willekeurig door elkaar. Tenslotte kan je uitschieters detecteren als de observaties met een gestandardizeerd residue dat in absolute waarde groter is dan 2.5.
- In de laatste figuur, tenslotte, zijn de predicties $\hat{y}_i$ uitgezet ten opzichte van de responsvariabele y_i. Als het model goed is, verwacht je dat de punten in de buurt van de eerste bissectrice liggen.

Stap 4: besluit

Nu we in stap 3 de modelveronderstellingen (de residuen zijn onafhankelijk en normaal verdeeld met dezelfde variantie) nagegaan en goedgekeurd hebben, kunnen we eindelijk een besluit formuleren.

.....

.....

2.2.2 Extra oefeningen

Oefening 5 We halen de dataset *zuurstof.xls* van Blackboard en importeren deze in R. Deze bevat informatie over zuurstof geproduceerd in een chemisch destillatieproces. Men heeft de zuiverheid van de zuurstof gemeten en de hoeveelheid koolwaterstof in de hoofdcondensator. De dataset bevat dan ook twee variabelen: *koolwaterstof* en *zuiverheid*.

1. Maak een scatterplot van de data. Denk je dat er een lineair verband te ontdekken valt?

   .....

2. Onderzoek de normaliteit van de gegevens. Let op: dit is geen voorwaarde voor regressie, wel voor de correlatietest.

   .....

   .....

3. Bereken een gepaste correlatiecoëfficiënt en de p-waarde van de bijbehorende test. Is het zinvol om een lineair verband te veronderstellen?

   .....

   .....

4. Wat is de vergelijking van de regressierechte en welke waarde heeft R^2?

   .....

5. Welke modelveronderstellingen heb je gemaakt? Ga deze ook na.

   .....

   .....

6. Wat is je algemeen besluit?

   .....

7. Welke waarde voor *koolwaterstof* verwacht je als de *zuiverheid* 91 zou bedragen?

   .....

Oefening 6 Open de dataset *politie.xls*. Deze bevat twee variabelen. De eerste (*agent*) bevat de verandering van het aantal politieagenten in New York (procentueel) en de tweede variabele (*diefstal*) geeft de verandering weer van het aantal diefstallen.

1. Onderzoek de normaliteit van de variabelen *agent* en *diefstal*.

   .....

.....

2. Bereken een gepaste correlatiecoëfficiënt en de p-waarde.

   .....

3. Wat kan je besluiten? Controleer je besluit aan de hand van een scatterplot.

   .....

4. Is het zinvol om een lineaire regressie uit te voeren?

   .....

5. Herhaal vraag 1 t.e.m. vraag 4 maar nu zonder de uitschieter in de variabele *agent*.

   .....

   .....

   .....

   .....