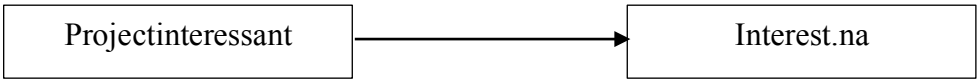
# Statistiek B – C4 - RESPONS

**Voorbereidend werk**

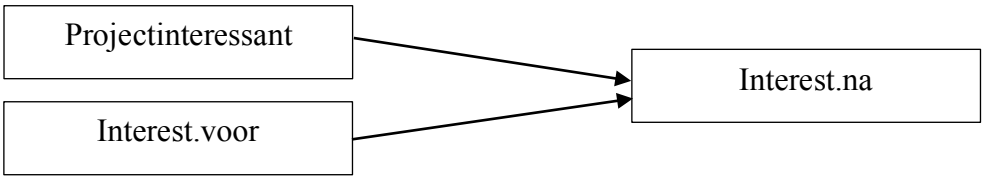
a) Teken de modellen.
b) Om alles eenvoudig interpreteerbaar te houden, maak je van alle variabelen die je nodig hebt eerst een z-score: 'Interest.na','Interest.voor', 'Projectinteressant', 'Projectleuk', 'Projectbijgeleerd', 'Projectmoeilijk'.
c) Om de verschillende modellen te kunnen vergelijken, maken we meteen gebruik van een databestand zonder 'NA'-waarden voor alle variabelen die je nodig hebt. (`functie: na.omit()`)

**VOORBEREIDEND WERK a**

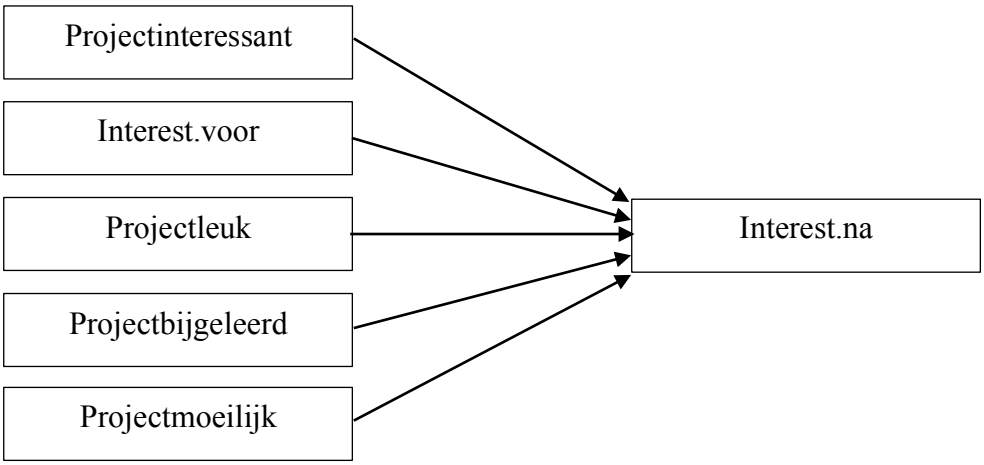
Model 1:




Model 2:




Model 3:



## VOORBEREIDEND WERK b en c

```
> Techniek$Interest.voorZ <- scale(Techniek$Interest.voor)
> Techniek$Interest.naZ <- scale(Techniek$Interest.na)
> Techniek$ProjectinteressantZ <- scale(Techniek$Projectinteressant)
> Techniek$ProjectmoeilijkZ <- scale(Techniek$Projectmoeilijk)
> Techniek$ProjectleukZ <- scale(Techniek$Projectleuk)
> Techniek$ProjectbijgeleerdZ <- scale(Techniek$Projectbijgeleerd)
> # variabelen standaardiseren, kan ook met functie z-score()
> DataC4 <- na.omit(Techniek[, c("Interest.voorZ", "Interest.naZ",
                                 "ProjectinteressantZ", "ProjectmoeilijkZ",
                                 "ProjectleukZ", "ProjectbijgeleerdZ")])
> # databestand aanmaken met volledige cases voor alle variabelen
```

**Oefening 1**

In een eerste model (Model1) onderzoeken we in welke mate het interessant vinden van het project ('Projectinteressant') een invloed heeft op de interesse in techniek na het project ('Interest.na').

a) Schat het model en bespreek de relevante parameters.
b) Ga de assumpties m.b.t. dit model na.

## OEFENING 1 a

```
> Model1 <- lm(Interest.naZ~ProjectinteressantZ, data = DataC4)
> summary(Model1)

Call:
lm(formula = Interest.naZ ~ ProjectinteressantZ, data = DataC4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4811 -0.6902 -0.0363  0.6031  2.6525

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.004071   0.019856  -0.205    0.838
ProjectinteressantZ  0.324501   0.019861  16.339   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9502 on 2288 degrees of freedom
Multiple R-squared:  0.1045,  Adjusted R-squared:  0.1041
F-statistic:   267 on 1 and 2288 DF,  p-value: < 2.2e-16
```

→ *$R^2$ = 0.10: het gaat om een medium effect (10% verklaarde variantie in 'Interest.naZ')*
*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*
*Dus we verwachten dat dit model in de populatie WEL variantie verklaart in 'Interest.naZ'.*
→ *intercept = -0.004: een leerling die 0 scoort op 'ProjectinteressantZ' scoort -0.004 op 'Interest.naZ'*
*Met p > 0.05: kans dat $H_0$ opgaat in de populatie is groter dan 5%*
*Dus we verwachten dit NIET in de populatie terug te vinden. Het verwachte intercept in de populatie is dus 0.*
*Dit is niet verwonderlijk aangezien zowel de onafhankelijke ('ProjectinteressantZ') als de afhankelijke variabele ('Interest.naZ') zijn gestandaardiseerd. Het intercept geeft hier dus de score weer op 'Interest.naZ' voor een leerling die gemiddeld scoort op*

*'ProjectinteressantZ'. M.a.w., in de populatie scoren leerlingen die gemiddeld scoren op 'ProjectinteressantZ' ook gemiddeld op 'Interest.naZ'.*

*→ $ß_{ProjectinteressantZ}$ = 0.32, dus 1 SD (in standaarddeviaties uitgedrukt want z-score) hoger scoren op 'ProjectinteressantZ' leidt tot 0.32 SD (in standaarddeviaties uitgedrukt want z-score) hoger scoren op 'Interest.naZ'*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

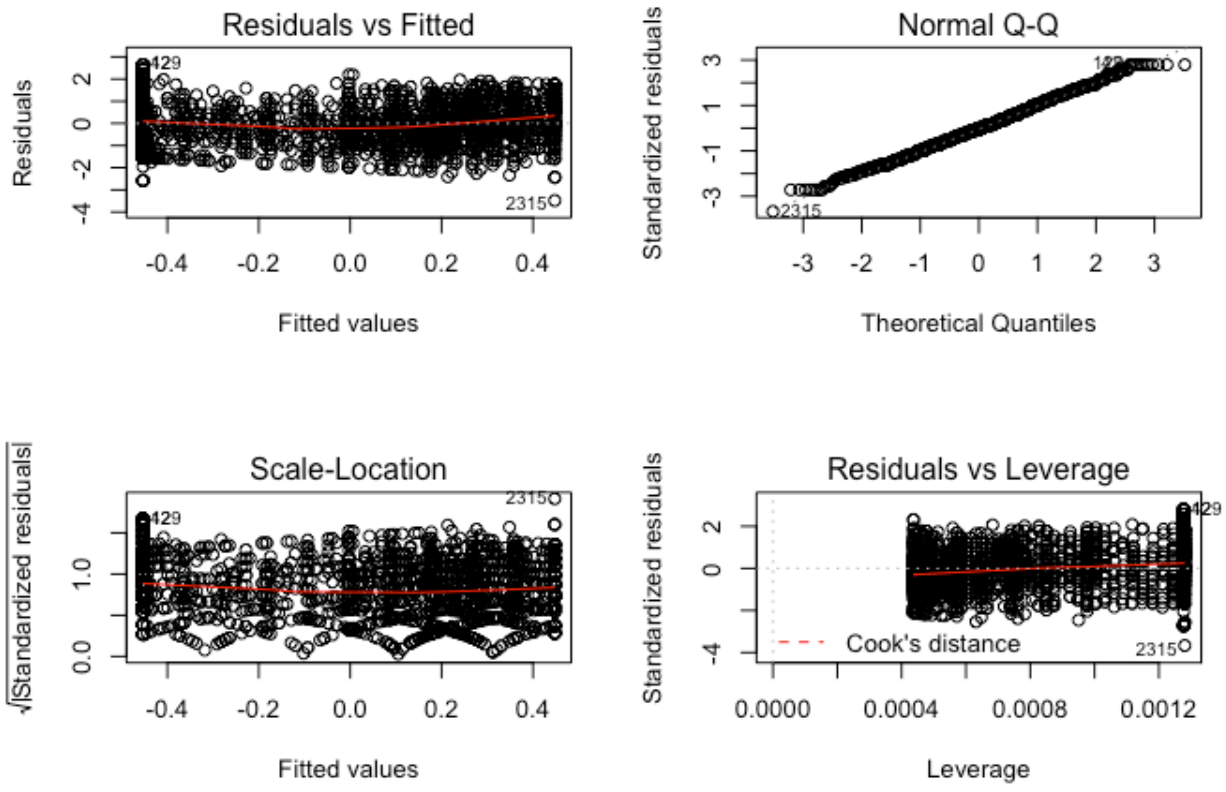
*Dus we verwachten dat 'ProjectinteressantZ' in de populatie WEL invloed heeft op 'Interest.naZ'.*

**CONCLUSIE:**

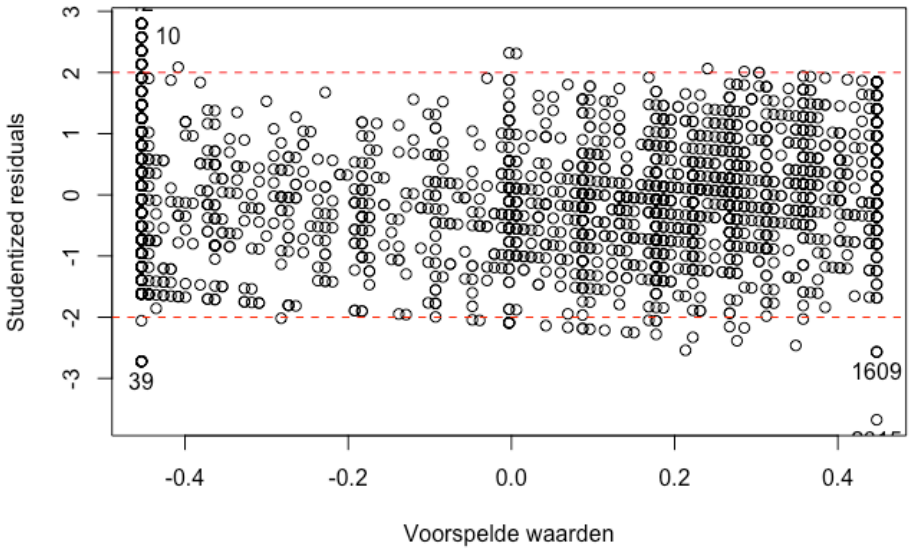
**De mate waarin leerlingen het project interessant vinden ('ProjectinteressantZ'), heeft een medium significant effect op hun interesse in techniek na het project ('Interest.naZ'; $R^2$ = 0.10; *p* < 0.05). Het effect is hier statistisch significant en positief (ß = 0.32, p < 0.05). Voor elke *SD* dat een leerling meer scoort op 'ProjectinteressantZ' zal de interesse in techniek na het project ('Interest.naZ') met 0.32 *SD* toenemen. Een leerling die 0 scoort op 'ProjectinteressantZ' scoort -0.004 op 'Interest.naZ'. Dit kunnen we echter niet doortrekken naar de populatie (p > 0.05). In de populatie verwachten we dat een leerling die het project gemiddeld interessant vindt, ook een gemiddelde interesse in techniek na het project zal hebben.**

**OEFENING 1 b**

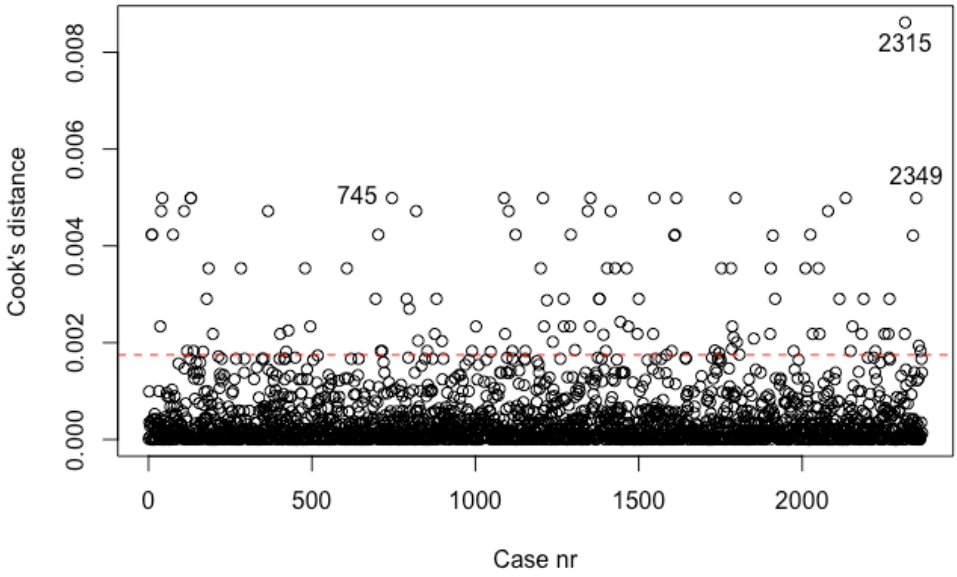
```
> par(mfrow = c(2, 2))
> plot(Model1)
```



```
> par(mfrow = c(1, 1)) #zo zet je het plotvenster terug op standaardweergave
> residuals_plot(Model1)
```




```
> cooks_plot(Model1)
```




→ *homoscedasticiteit:*

*Grafiek 'Residuals versus Fitted': spreiding blijft quasi gelijk naarmate de fitted value toeneemt. Dit duidt op homoscedasticiteit.*

*Grafiek 'Scale-location plot': de rode lijn in de plot loopt min of meer recht. Dit duidt op homoscedasticiteit.*

→ *errortermen normaal verdeeld:*

*Grafiek 'Q-Q plot': alle punten vallen min of meer op de rechte lijn. Dit geeft aan dat de errortermen normaal verdeeld zijn.*

→ *geen clustering en een lineair verband:*

*Grafiek 'Residuals versus Fitted': op deze grafiek is geen clustering van waarnemingen terug te vinden. Noch kan je op basis van deze grafiek een niet-lineair verband vermoeden.*

→ *wel outliers:*

*Grafieken 'Residuals plot' & 'Cooks plot': duidelijk aantal outliers zichtbaar. Deze zouden verwijderd moeten worden om vervolgens het model opnieuw te schatten en opnieuw te assumpties na te gaan.*

**CONCLUSIE:**
**Op basis van deze grafieken kunnen we ervan uitgaan dat bijna aan alle assumpties m.b.t. regressieanalyse is voldaan. Er is enkel sprake van enkele outliers.**

**Oefening 2**

Wanneer blijkt dat de mate waarin de leerling het project als interessant ervaart er toe doet, kun je je natuurlijk afvragen hoe dat komt. Misschien is het zo dat leerlingen die techniek sowieso al interessant vonden voor het project ('Interest.voor') nadien ook een hogere interesse behouden (die hadden ze tenslotte al voor het project). Om dit na te gaan, test je een tweede model (Model2) waarin je 'Interest.voor' als controlevariabele aan het vorige model (Model1) toevoegt.

a) Is dit model (Model2) een beter model dan Model1?
b) Bespreek de relevante parameters van het beste model.

**OEFENING 2 a**

```
> Model2 <- lm(Interest.naZ ~ ProjectinteressantZ + Interest.voorZ, data =
DataC4)
> anova(Model1, Model2)
Analysis of Variance Table

Model 1: Interest.naZ ~ ProjectinteressantZ
Model 2: Interest.naZ ~ ProjectinteressantZ + Interest.voorZ
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2288 2065.7
2   2287 1310.5  1    755.26 1318.1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ *$RSS_{Model1}$ = 2065.7, $RSS_{Model2}$ = 1310.5, $p < 0.05$*

*Model2 is statistisch significant beter dan Model1. Model2 heeft een lagere RSS en het verschil in RSS (ΔRSS = 755.26) is statistisch significant ($p < 0.05$). Dus we verwachten dit verschil in RSS WEL in de populatie.*

**OEFENING 2 b**

```
> summary(Model2)

Call:
lm(formula = Interest.naZ ~ ProjectinteressantZ + Interest.voorZ,
    data = DataC4)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0384 -0.4520  0.0093  0.4924  2.7215
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.003766   0.015818  -0.238    0.812
ProjectinteressantZ  0.180850   0.016310  11.089   <2e-16 ***
Interest.voorZ       0.592104   0.016309  36.305   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.757 on 2287 degrees of freedom
Multiple R-squared:  0.4319,  Adjusted R-squared:  0.4314
F-statistic: 869.3 on 2 and 2287 DF,  p-value: < 2.2e-16
```

→ *$R^2$ = 0.43: het gaat om een groot effect (43% verklaarde variantie in 'Interest.naZ')*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

*Dus we verwachten dat dit model in de populatie WEL variantie verklaart in 'Interest.naZ'.*

→ *intercept = -0.004: een leerling die 0 scoort op 'ProjectinteressantZ' en 'Interest.voor' scoort -0.004 op 'Interest.naZ'*

*Met p > 0.05: kans dat $H_0$ opgaat in de populatie is groter dan 5%*

*Dus we verwachten dit NIET in de populatie terug te vinden. Het verwachte intercept in de populatie is dus 0. Dit is niet verwonderlijke aangezien zowel de onafhankelijke variabelen ('Interest.voorZ' en 'ProjectinteressantZ') als de afhankelijke variabele ('Interest.naZ') zijn gestandaardiseerd. Het intercept geeft hier dus de score weer op 'Interest.naZ' voor leerlingen die gemiddeld scoren op 'Interest.voorZ' en 'ProjectinteressantZ'. M.a.w., in de populatie scoren leerlingen die gemiddeld scoren op 'Interest.voorZ' en 'ProjectinteressantZ' ook gemiddeld op 'Interessant.na'.*

→ *$ß_{ProjectinteressantZ}$ = 0.18, dus 1 SD (want z-score!) hoger scoren op 'Projectinteressant' leidt tot 0.18 SD (want z-score!) hoger scoren op 'Interest.na'*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

*Dus we verwachten dat 'Projectinteressant' in de populatie WEL invloed heeft op 'Interest.na'.*

→ *$ß_{Interest.voorZ}$ = 0.59, dus 1 SD (want z-score!) hoger scoren op 'Interest.voorZ' leidt tot 0.59 SD (want z-score!) hoger scoren op 'Interest.na'*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

*Dus we verwachten dat 'Interest.voorZ' in de populatie WEL invloed heeft op 'Interest.naZ'. Bovendien is dit effect sterker dan dat van 'ProjectinteressantZ'. (Je mag de sterkte van beide effecten met elkaar vergelijken, omdat beide variabelen gestandaardiseerd zijn en dus op dezelfde schaal staan.)*

**<u>CONCLUSIE:</u>**

**Interesse in techniek voor het project ('Interest.voorZ') en de mate waarin leerlingen het project interessant ('ProjectinteressantZ') vinden verklaren de interesse in techniek van leerlingen na het project ('Interest.naZ') beter als het model waarin enkel 'ProjectinteressantZ' als onafhankelijke variabele is opgenomen (*$\Delta RSS$* = 755.26, *$p$* < 0.05). Het gaat bovendien om een sterk, significant effect (*$R^2$* = 0.43, *$p$* < 0.05).**

**Het intercept is niet statistisch significant (*$p$* > 0.05). Een leerling die gemiddeld scoort op 'ProjectinteressantZ' en 'Interest.voorZ' scoort dus ook gemiddeld op 'Interest.naZ' in de populatie. (Wat logisch is, aangezien het hier om gestandaardiseerde variabelen gaat.)**

**Zowel 'ProjectinteressantZ' als 'Interest.voorZ' hebben een positief en statistisch significant (*$p$* < 0.05) effect op 'Interest.naZ'. Een toename van 1 *SD* in 'ProjectinteressantZ' leidt tot een toename van 0.18 *SD* in 'Interest.naZ'. Het effect van 'Interest.voorZ' is sterker (*ß* = 0.59). 1 *SD* hoger scoren op interesse in techniek voor het project leidt tot een toename van 0.59 *SD* in interesse in techniek na het project.**

**Oefening 3**

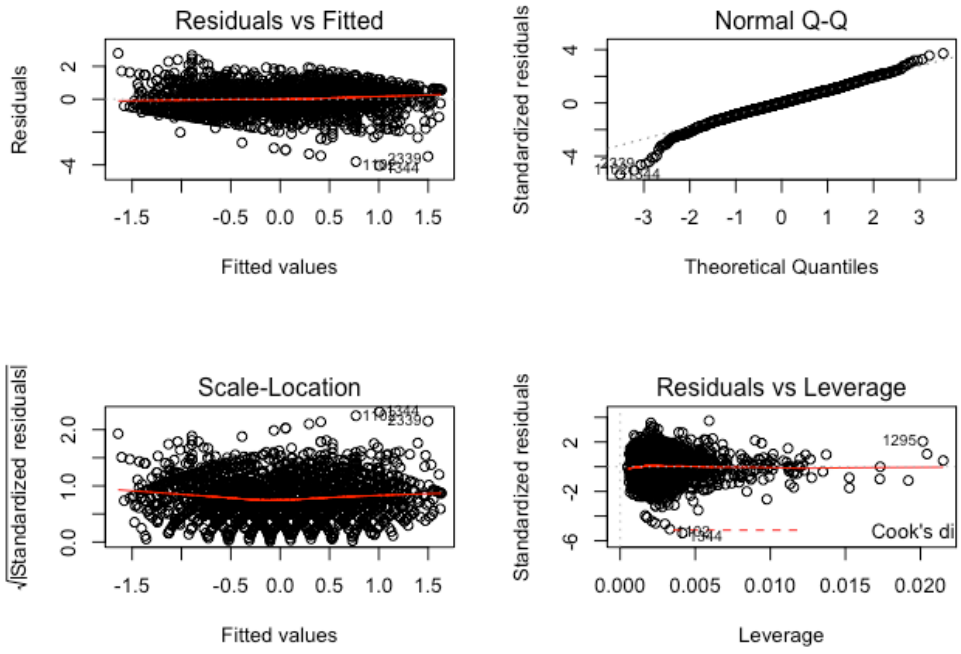
Mogelijk zorgt het opnemen van de overige variabelen die de houdingen van de leerlingen m.b.t. het techniekproject meten ('Projectleuk', 'Projectbijgeleerd', 'Projectmoeilijk') voor een verbetering van het model. Om dit na te gaan, test je een derde model (Model3) waarin je deze drie variabelen als controlevariabelen aan het vorige model (Model2) toevoegt.

a) Gaan de assumpties op? Pas het model indien nodig aan.
b) Is Model3 een beter model dan Model2?
c) Bespreek de relevante parameters van het beste model.
d) Welke score op 'Interest.naZ' behaalt iemand die:
    a. gemiddeld scoort op alle onafhankelijke variabelen *in de steekproef*
    b. gemiddeld scoort op alle onafhankelijke variabelen *in de populatie*
    c. 1 *SD* hoger scoort voor 'Interest.voorZ' en 1 *SD* lager scoort voor 'ProjectbijgeleerdZ' (en op alle andere onafhankelijke variabelen 0) *in de steekproef*
    d. 1 *SD* hoger scoort voor 'Interest.voorZ' en en 1 *SD* lager scoort voor 'ProjectbijgeleerdZ' (en op alle andere onafhankelijke variabelen 0) *in de populatie*

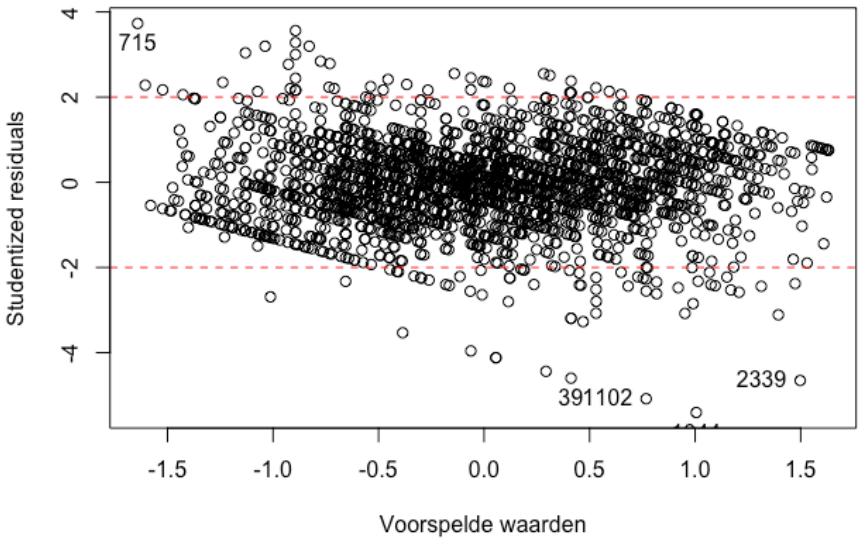
**Rond bij het berekenen van de voorspelde scores altijd af tot op 3 cijfers na de komma!**

## OEFENING 3 a

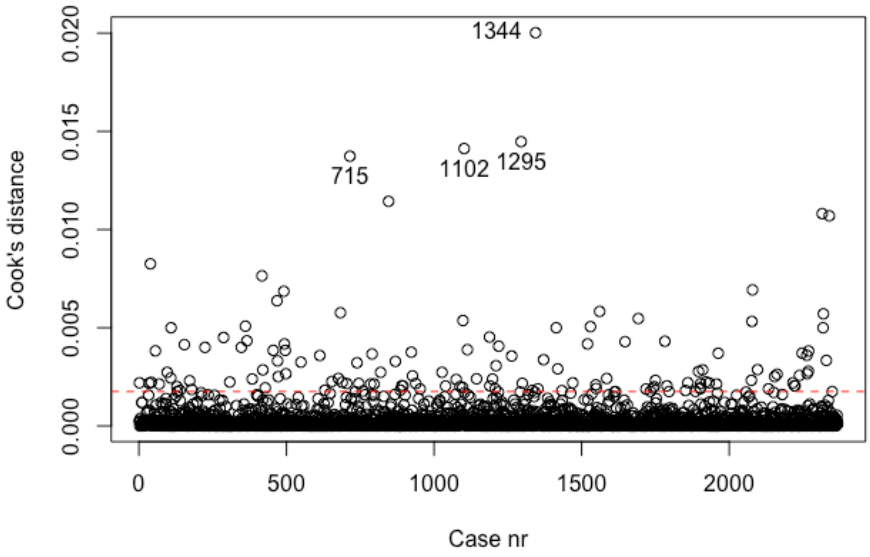
```
> Model3 <- lm(Interest.naZ ~ ProjectinteressantZ + Interest.voorZ +
+                             ProjectleukZ + ProjectbijgeleerdZ +
+                             ProjectmoeilijkZ, data = DataC4)

> par(mfrow = c(2,2))
> plot(Model3)
```




```
> par(mfrow = c(1,1))
> residuals_plot(Model3)
```




```
> cooks_plot(Model3)
```




```
> library(car)
> vif(Model3)
ProjectinteressantZ      Interest.voorZ         ProjectleukZ
          7.041215           1.117586           5.554361
 ProjectbijgeleerdZ     ProjectmoeilijkZ
          3.859391           1.278241
> cor.test(DataC4$ProjectinteressantZ, DataC4$ProjectleukZ)

        Pearson's product-moment correlation

data:  DataC4$ProjectinteressantZ and DataC4$ProjectleukZ
t = 99.142, df = 2288, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8926227 0.9081126
sample estimates:
      cor
0.9006534
```

→ *homoscedasticiteit:*

*Grafiek 'Residuals versus Fitted': spreiding blijft quasi gelijk naarmate de fitted value toeneemt. Dit duidt op homoscedasticiteit.*

*Grafiek 'Scale-location plot': de rode lijn in de plot loopt min of meer recht. Dit duidt op homoscedasticiteit.*

→ *errortermen niet-normaal verdeeld:*

*Grafiek 'Q-Q plot': niet alle punten vallen min of meer op de rechte lijn. Dit geeft aan dat de errortermen mogelijk niet normaal verdeeld zijn.*

→ *geen clustering en een lineair verband:*

*Grafiek 'Residuals versus Fitted': op deze grafiek is geen clustering van waarnemingen terug te vinden. Noch kan je op basis van deze grafiek een niet-lineair verband vermoeden.*

→ *wel outliers:*

*Grafieken 'Residuals plot & 'cooks plot': duidelijk aantal outliers zichtbaar. Deze zouden verwijderd moeten worden om vervolgens het model opnieuw te schatten en opnieuw te assumpties af te toetsen.*

→ *wel multicollineariteit:*

*Uit de VIF-waarden blijkt dat er mogelijk een probleem is met de samenhang tussen de variabelen 'ProjectinteressantZ' en 'ProjectleukZ'. Deze variabelen blijken zeer sterk te correleren (r = 0.90). Daarom is het aangewezen om het model de herdraaien zonder één van beide variabelen. Anders levert dit vertekende significantietoetsen voor de betrokken variabelen op. Daarom herschatten we het model zonder de controlevariabele 'ProjectleukZ'. Uit de output hieronder blijkt dat het multicollineariteitsprobleem dan is opgelost. We werken dus verder met Model4.*

```
> Model4 <- lm(Interest.naZ ~ ProjectinteressantZ + Interest.voorZ +
+                             ProjectbijgeleerdZ + ProjectmoeilijkZ,
+                             data = DataC4)
> par(mfrow=c(2,2))
> plot(Model4)
> vif(Model4)
ProjectinteressantZ      Interest.voorZ  ProjectbijgeleerdZ
          3.671585            1.107204            3.725005
   ProjectmoeilijkZ
          1.278237
```

**CONCLUSIE:**

**Op basis van deze grafieken kunnen we ervan uitgaan dat mogelijk niet aan de assumpties m.b.t. regressieanalyse is voldaan. Naast enkele outliers en het mogelijke probleem met de niet-normaal verdeelde errortermen vraagt voornamelijk het multicollineariteitsprobleem om het herschatten van Model3 zonder de variabele 'ProjectleukZ'.**

## OEFENING 3 b

```
> anova(Model2, Model4)
Analysis of Variance Table

Model 1: Interest.naZ ~ ProjectinteressantZ + Interest.voorZ
Model 2: Interest.naZ ~ ProjectinteressantZ + Interest.voorZ +
ProjectbijgeleerdZ +
    ProjectmoeilijkZ
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2287 1310.5
2   2285 1297.3  2    13.166 11.595 9.765e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ *$RSS_{Model2}$ = 1310.5, $RSS_{Model4}$ = 1297.3, p < 0.05*

*Model4 is statistisch significant beter dan Model2. Model4 heeft een lagere RSS. Het verschil in RSS (ΔRSS = 13.17) is statistisch significant (p < 0.05). Dus we verwachten dit verschil in RSS WEL in de populatie.*

**CONCLUSIE:**
**Model4 verklaart de verschillen in interesse in techniek na het project ('Interest.naZ') beter als Model2 (*ΔRSS* = 13.17, *p* < 0.05).**

## OEFENING 3 c

```
> summary(Model4)

Call:
lm(formula = Interest.naZ ~ ProjectinteressantZ + Interest.voorZ +
    ProjectbijgeleerdZ + ProjectmoeilijkZ, data = DataC4)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0431 -0.4473  0.0046  0.4895  2.7835

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.003901   0.015746  -0.248    0.804
ProjectinteressantZ  0.188626   0.030178   6.250 4.87e-10 ***
Interest.voorZ       0.576293   0.016572  34.776  < 2e-16 ***
ProjectbijgeleerdZ   0.034290   0.030389   1.128    0.259
ProjectmoeilijkZ    -0.085936   0.017852  -4.814 1.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7535 on 2285 degrees of freedom
Multiple R-squared:  0.4376,  Adjusted R-squared:  0.4366
F-statistic: 444.5 on 4 and 2285 DF,  p-value: < 2.2e-16
```

→ *$R^2$ = 0.44: het gaat om een groot effect (44% verklaarde variantie in 'Interest.naZ')*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

*Dus we verwachten dat dit model in de populatie WEL variantie verklaart in 'Interest.naZ'.*

→ *intercept = -0.004: een leerling die 0 scoort op 'ProjectinteressantZ', 'Interest.voor', 'ProjectbijgeleerdZ' en 'ProjectmoeilijkZ' behaalt -0.004 op 'Interest.naZ'*

*Met p > 0.05: kans dat $H_0$ opgaat in de populatie is groter dan 5%*

*Dus we verwachten dit NIET in de populatie terug te vinden. Het verwachte intercept in de populatie is dus 0. Dit is niet verwonderlijke aangezien zowel de onafhankelijke*

*variabelen ('Interest.voorZ', 'ProjectinteressantZ', 'ProjectbijgeleerdZ', 'ProjectmoeilijkZ') als de afhankelijke variabele ('Interest.naZ') zijn gestandaardiseerd. Het intercept geeft hier dus de score weer op 'Interest.naZ' voor leerlingen die gemiddeld scoren op alle onafhankelijke variabelen. M.a.w., in de populatie scoren leerlingen die gemiddeld scoren op 'Interest.voorZ', 'ProjectinteressantZ', 'ProjectmoeilijkZ' en 'ProjectbijgeleerdZ' ook gemiddeld op 'Interest.naZ'.*

→ *$ß_{ProjectinteressantZ}$ = 0.19, dus 1 SD (want z-score!) hoger scoren op 'ProjectinteressantZ' leidt tot 0.19 SD (want z-score!) hoger scoren op 'Interest.naZ'*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

*Dus we verwachten dat 'ProjectinteressantZ' in de populatie WEL invloed heeft op 'Interest.naZ'.*

→ *$ß_{Interest.voorZ}$ = 0.58, dus 1 SD (want z-score!) hoger scoren op 'Interest.voorZ' leidt tot 0.58 SD (want z-score!) hoger scoren op 'Interest.naZ'*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

*Dus we verwachten dat 'Interest.voorZ' in de populatie WEL invloed heeft op 'Interest.naZ'. Bovendien is dit effect sterker dan dat van 'ProjectinteressantZ' en dat van 'ProjectmoeilijkZ'. (Je mag de sterkte van deze effecten met elkaar vergelijken, omdat beide variabelen gestandaardiseerd zijn en dus op dezelfde schaal staan.)*

→ *$ß_{ProjectibijgeleerdZ}$ = 0.03, dus 1 SD (want z-score!) hoger scoren op 'ProjectbijgeleerdZ' leidt tot 0.03 SD (want z-score!) hoger scoren op 'Interest.naZ'*

*Met p > 0.05: kans dat $H_0$ opgaat in de populatie is groter dan 5%*

*Dus we verwachten dat 'ProjectinteressantZ' in de populatie GEEN invloed heeft op 'Interest.naZ'.*

→ *$ß_{ProjectmoeilijkZ}$ = -0.09, dus 1 SD (want z-score!) hoger scoren op 'ProjectmoeilijkZ' leidt tot 0.09 SD (want z-score!) lager scoren op 'Interest.naZ'*

*Met p < 0.05: kans dat $H_0$ opgaat in de populatie is kleiner dan 5%*

*Dus we verwachten dat 'ProjectmoeilijkZ' in de populatie WEL invloed heeft op 'Interest.naZ'.*

**CONCLUSIE:**

**'ProjectinteressantZ', 'Interest.voorZ', 'ProjectbijgeleerdZ' en 'ProjectmoeilijkZ' verklaren 43% van de variantie in 'Interest.naZ'. Het gaat dus om een sterk effect dat bovendien mag worden doorgetrokken naar de populatie ($R^2$ = 0.44, $p$ < 0.05).**

**Het intercept is niet statistisch significant ($p$ > 0.05). Een leerling die gemiddeld scoort op 'ProjectinteressantZ', 'Interest.voorZ', 'ProjectbijgeleerdZ' en 'ProjectmoeilijkZ' scoort dus ook gemiddeld op 'Interest.naZ' in de populatie. (Wat logisch is, aangezien het hier om gestandaardiseerde variabelen gaat.)**

**Zowel 'ProjectinteressantZ' als 'Interest.voorZ' hebben een positief en statistisch significant ($p$ < 0.05) effect op 'Interest.naZ'. Een toename van 1 *SD* in 'ProjectinteressantZ' leidt tot een toename van 0.19 *SD* in 'Interest.naZ'. Het effect van 'Interest.voorZ' is sterker ($ß$ = 0.58). 1 *SD* hoger scoren op interesse in techniek voor het project ('Interest.voorZ') leidt tot een toename van 0.58 *SD* in interesse in techniek na het project ('Interest.naZ'). De invloed van 'ProjectmoeilijkZ' op 'Interest.naZ' is negatief en statistisch significant (p < 0.05). Leerlingen die het project als 1 *SD* moeilijker ervaren, scoren 0.09 *SD* lager op 'Interest.naZ'. De percepties van leerlingen m.b.t. bijleren tijdens het project ('ProjectbijgeleerdZ') hebben geen significante invloed op hun interesse in techniek na het project (p > 0.05).**

**OEFENING 3 d**

Om onderstaande scores te berekenen, gebruiken we de output van Model4 en de regressievergelijking.

Output model 4:

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         -0.003901   0.015746  -0.248    0.804
ProjectinteressantZ  0.188626   0.030178   6.250 4.87e-10 ***
Interest.voorZ       0.576293   0.016572  34.776  < 2e-16 ***
ProjectbijgeleerdZ   0.034290   0.030389   1.128    0.259
ProjectmoeilijkZ    -0.085936   0.017852  -4.814 1.58e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Regressievergelijking:

Interest.naZ = Intercept + $ß_1$*ProjectinteressantZ + $ß_2$*Interest.voorZ + $ß_3$*ProjectbijgeleerdZ + $ß_4$*ProjectmoeilijkZ

Welke score op 'Interest.naZ' behaalt iemand die:

a) gemiddeld scoort op alle onafhankelijke variabelen *in de steekproef:*

Interest.naZ = -0.004 + 0*0.189 + 0*0.576 + 0*0.034 + 0*-0.086
= -0.004 (= het intercept uit de steekproef!!!)

b) gemiddeld scoort op alle onafhankelijke variabelen *in de populatie:*

Interest.naZ = -0.004 + 0*0.189 + 0*0.576 + 0*0 + 0*-0.086
= 0 (= het intercept in de populatie!!!)

In de populatie bedraagt het verwachte intercept 0, aangezien het NIET statistisch significant is ($p > 0.05$).

c) 1 *SD* hoger scoort voor 'Interest.voorZ' en 1 *SD* lager scoort voor 'ProjectbijgeleerdZ' (en op alle andere onafhankelijke variabelen 0) *in de steekproef*

Interest.naZ = -0.004 + 0*0.189 + 1*0.576 - 1*0.034 + 0*-0.086
= -0.004 + 0.576 – 0.034
= 0.546

d) 1 *SD* hoger scoort voor 'Interest.voorZ' en 1 *SD* lager scoort voor 'ProjectbijgeleerdZ' (en op alle andere onafhankelijke variabelen 0) *in de populatie*

Interest.naZ = -0.004 + 0*0.189 + 1*0.576 - 1*0 + 0*-0.086
= 0 + 0.576
= 0,58

In de populatie valt het effect van 'ProjectbijgeleerdZ' weg (= 0), aangezien het NIET statistisch significant is ($p > 0.05$).