

Contactmoment 6: Respons computer lab

! Important

Vooraleer je de oefeningen kan oplossen is het belangrijk om zowel de dataset te laden, het pakket `car` te activeren en ook de OLP2 Functies te activeren.

Oefening 1

Oefening 1

Is er een verschil tussen eerste en tweedejaars in technische geletterdheid (`TAC.na`) en is het zo dat het effect van “eigen inbreng in de les” (`PISA_EigenInbreng`) op de technische geletterdheid anders is voor eerste dan wel voor tweedejaars (`Studiejaar`)?

Vooraleer je deze analyse draait, maak je een dummyvariabele aan voor de variabele `Studiejaar`. Zorg ervoor dat deze dummy aanstaat voor de categorie “tweedejaars”. Hanteer deze dummyvariabele in je verdere analyses.

We starten met het aanmaken van de dummy variabele en checken of we dit goed gedaan hebben door een tabel aan te maken waarin we de oorspronkelijke variabele kruisen met de aangemaakte dummy variabele.

```
Techniek$Tweedejaars<-(Techniek$Studiejaar=='2')*1  
table(Techniek$Studiejaar,Techniek$Tweedejaars)
```

	0	1
1	1083	0
2	0	1284

We schatten nu het model met daarin enerzijds de hoofdeffecten van `PISA_EigenInbreng` en `Tweedejaars` en anderzijds het interactie-effect tussen beide. (Je merkt in onderstaand

codeblok dat we werken met wat meer ‘enters’ en de code over lijnen spreiden om ze leesbaar te houden)

```
M_EigenInbreng <- lm(
  TAC.na ~ PISA_EigenInbreng + Tweedejaars + PISA_EigenInbreng*Tweedejaars,
  data=Techniek
)

summary(M_EigenInbreng)
```

Call:

```
lm(formula = TAC.na ~ PISA_EigenInbreng + Tweedejaars + PISA_EigenInbreng *
    Tweedejaars, data = Techniek)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.62846	-0.14602	0.01189	0.16344	0.47802

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.679370	0.018249	37.228	< 2e-16 ***
PISA_EigenInbreng	-0.053860	0.009498	-5.671	1.67e-08 ***
Tweedejaars	0.010221	0.024576	0.416	0.678
PISA_EigenInbreng:Tweedejaars	-0.007273	0.013114	-0.555	0.579

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2041 on 1675 degrees of freedom
(688 observations deleted due to missingness)

Multiple R-squared: 0.04447, Adjusted R-squared: 0.04275

F-statistic: 25.98 on 3 and 1675 DF, p-value: < 2.2e-16

De mate waarin leerlingen eigen inbreng mogen doen, het studiejaar waarin leerlingen zitten en de interactie tussen deze beide variabelen verklaren samen 4.28% van de variantie ($R^2 = 0.0428$) in technische geletterdheid (TAC.na). We verwachten dat deze variabelen ook in de populatie een invloed gaan hebben op technische geletterdheid ($F(3,1675) = 25.98$, $p < 0.001$).

Aangaande de vraag of er een verschil is in technische geletterdheid tussen eerste- en tweedejaars studenten stellen we vast dat er geen statistisch significant effect is van de variabele *Tweedejaars* op TAC.na ($p = 0.01$, $p = 0.678$).

De variabele `PISA_EigenInbreng` heeft wel een statistisch significant hoofdeffect ($\beta = -0.05$, $p < 0.05$). Leerlingen die één eenheid hoger scoren op de mate waarin ze vinden eigen inbreng te mogen hebben in de les scoren 0.05 eenheden lager op technische geletterdheid. In de populatie verwachten we ook een effect gegeven de p-waarde die kleiner is dan 0.05. Onze hypothese met betrekking tot verschil in het effect van `PISA_EigenInbreng` tussen eerste- en tweedejaars dienen we te verwerpen. Het is niet zo dat leerlingen in het tweede jaar meer baat hebben bij het hebben van eigen inbreng in de les. Het interactie-effect van deze lesstijl met het jaar is niet significant ($\beta = -0.007$, $p = 0.579$). Anders gesteld: leerlingen uit het eerste of tweede jaar hebben dezelfde mate van eigen inbreng nodig in de les om technisch geletterder te worden.

Oefening 2

Oefening 2

En hoe zit dat met de andere leerkrachtstijlen (`PISA_Experimenteren`; `PISA_Orientatie`)? Hebben die hetzelfde effect voor eerste en tweedejaars? Gebruik opnieuw de dummyvariabele (`Tweedejaars`) om deze hypothesen te toetsen. Toets elk van die hypothesen in een afzonderlijk model.

We gaan eerst de twee nieuwe modellen schatten, één met daarin `PISA_Experimenteren` als voorspeller en één met `PISA_Orientatie` als voorspeller.

```
M_Experimenteren <- lm(
  TAC.na ~ PISA_Experimenteren + Tweedejaars + PISA_Experimenteren * Tweedejaars,
  data = Techniek
)

M_Orientatie <- lm(
  TAC.na ~ PISA_Orientatie + Tweedejaars + PISA_Orientatie * Tweedejaars,
  data = Techniek
)
```

We bekijken de output voor beide modellen.

```
summary(M_Experimenteren)
```

Call:

```
lm(formula = TAC.na ~ PISA_Experimenteren + Tweedejaars + PISA_Experimenteren *
    Tweedejaars, data = Techniek)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.59012	-0.16213	0.00697	0.18210	0.41744

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.582276	0.029566	19.694	<2e-16 ***
PISA_Experimenteren	0.001171	0.012381	0.095	0.925
Tweedejaars	0.028006	0.038866	0.721	0.471
PISA_Experimenteren:Tweedejaars	-0.011252	0.016513	-0.681	0.496

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2088 on 1675 degrees of freedom

(688 observations deleted due to missingness)

Multiple R-squared: 0.000561, Adjusted R-squared: -0.001229

F-statistic: 0.3134 on 3 and 1675 DF, p-value: 0.8157

`summary(M_Orientatie)`

Call:

```
lm(formula = TAC.na ~ PISA_Orientatie + Tweedejaars + PISA_Orientatie *  
    Tweedejaars, data = Techniek)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.60970	-0.16449	0.00952	0.17363	0.42555

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.544560	0.028204	19.308	<2e-16 ***
PISA_Orientatie	0.016128	0.010834	1.489	0.137
Tweedejaars	-0.005354	0.035844	-0.149	0.881
PISA_Orientatie:Tweedejaars	0.005019	0.014177	0.354	0.723

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2083 on 1675 degrees of freedom

(688 observations deleted due to missingness)

Multiple R-squared: 0.004543, Adjusted R-squared: 0.002761

F-statistic: 2.548 on 3 and 1675 DF, p-value: 0.05431

Wat beide andere lesstijlen betreft, blijkt dat de verklaarde variantie in technische geletterdheid (**TAC.na**) telkens verwaarloosbaar is (R-kwadraat **Experimenteren**= -0.0012, R-kwadraat **Orientatie**= 0.0028) en bovendien niet statistisch significant (p **Experimenteren**= 0.816, p **Orientatie**= 0.054). De onafhankelijke variabelen hebben dus geen effect op technische geletterdheid in de populatie. Bovendien is voor beide modellen zowel **Tweedejaars** (**Experimenteren** = -0.007, p = 0.471; **Orientatie**= 0.016, p = 0.137) als de interactie tussen **Tweedejaars** en lesstijl (**ExperimenterenTweedejaars** = -0.011, p = 0.496; **OrientatieTweedejaars**= 0.005, p = 0.723) niet statistisch significant. In de populatie vinden we deze effecten dus niet terug.

Oefening 3

Oefening 3

Tijdens een vorig contactmoment gingen we al na in welke mate de interesse van leerlingen in het project (**Projectinteressant.z**) en de mate van interesse in techniek voor het project (**Interest.voor.z**) een invloed hebben op de interesse na het project (**Interest.na.z**).

Op basis van een uitgebreide literatuurstudie vermoeden we echter dat de interesse in techniek na het project niet voor alle studierichtingen hetzelfde is.

- (a) Ga deze hypothese na. Zet de variabele **Richting5cat** eerst om in de nodige dummyvariabelen. Hanteer vervolgens de studierichting “Technisch” als referentiecategorie in je analyses. Maak ook de nodige gestandaardiseerde variabelen aan.
- (b) Interpreteer vervolgens ook de andere parameterschattingen.

(a)

We starten met het aanmaken van de nodige dummy-variabelen.

Note

Om te weten welke code van **Richting5cat** hoort bij welke studierichting grijp je best terug naar het codeboek.

In de onderstaande code maken we de dummy-variabelen aan. Soms voeren we nog een extra controle uit via het **table()** commando. Dat doen we hier niet. We zijn overtuigd dat we het goed deden !

```
Techniek$Technisch <- (Techniek$Richting5cat=='1')*1
Techniek$Kunst     <- (Techniek$Richting5cat=='2')*1
```

```
Techniek$Latijn    <- (Techniek$Richting5cat=='3')*1
Techniek$ModWet    <- (Techniek$Richting5cat=='4')*1
Techniek$Handel    <- (Techniek$Richting5cat=='5')*1
```

Daarnaast moeten we enkele variabelen standaardiseren.

```
Techniek$Projectinteressant.z <- scale(Techniek$Projectinteressant)
Techniek$Interest.voor.z      <- scale(Techniek$Interest.voor)
Techniek$Interest.na.z       <- scale(Techniek$Interest.na)
```

Nu zijn we klaar om het gevraagde model te schatten en de output op te vragen.

```
Model3a<-lm(
  Interest.na.z ~ Projectinteressant.z +
    Interest.voor.z +
    Kunst +
    Latijn +
    ModWet +
    Handel,
  data=Techniek
)
summary(Model3a)
```

Call:

```
lm(formula = Interest.na.z ~ Projectinteressant.z + Interest.voor.z +
    Kunst + Latijn + ModWet + Handel, data = Techniek)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.9210	-0.4538	0.0149	0.4781	2.7577

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.28057	0.05886	4.767	1.99e-06 ***
Projectinteressant.z	0.18773	0.01643	11.423	< 2e-16 ***
Interest.voor.z	0.56743	0.01706	33.259	< 2e-16 ***
Kunst	-0.02814	0.13274	-0.212	0.832
Latijn	-0.33973	0.06588	-5.157	2.73e-07 ***
ModWet	-0.31074	0.06355	-4.890	1.08e-06 ***
Handel	-0.30135	0.07370	-4.089	4.49e-05 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7497 on 2236 degrees of freedom

(124 observations deleted due to missingness)

Multiple R-squared: 0.4403, Adjusted R-squared: 0.4388

F-statistic: 293.2 on 6 and 2236 DF, p-value: < 2.2e-16

De hypothese klopt! Uit de analyses blijkt dat leerlingen die gemiddeld scoren op `Projectinteressant.z` en `Interest.voor.z` maar les volgen in de studierichtingen Latijn ($= -0.34$, $p < 0.001$), Moderne Wetenschappen ($= -0.311$, $p < 0.001$) of Handel ($= -0.301$, $p < 0.001$) significant lager scoren op `Interest.na.z` in vergelijking met leerlingen uit een technische richting. Het verschil in score op `Interest.na.z` tussen leerlingen uit Kunst en leerlingen uit een technische richting is niet significant ($= -0.028$, $p = 0.832$).

(b)

Een leerling uit een technische richting die gemiddeld scoort op ‘`Projectinteressant.z`’ en op ‘`Interest.voor.z`’ behaalt een score van 0.281 SD boven het gemiddelde ($0 = 0.281$, $p < 0.001$). De kans dat het intercept in de populatie 0 bedraagt, is zeer klein ($p < 0.001$). We gaan er dus vanuit dat dit ook voor de populatie geldt. Als we het effect van `Projectinteressant.z` op `Interest.na.z` vergelijken met dat van `Interest.voor.z` op `Interest.na.z`, dan blijkt dat het effect van de laatste variabele groter is. We mogen beide effecten vergelijken, omdat het in beide gevallen om gestandaardiseerde variabelen gaat.

`Interest.voor.z` heeft een positief effect op `Interst.na.z` ($= 0.567$, $p < 0.001$). Voor elke SD die een leerling hoger scoort op `Interest.voor.z` stijgt diens score op `Interest.na.z` met 0.567 SD, ongeacht de studierichting waarin de leerling les volgt en diens score op `Project.interessant.z`. (Ook de toename in `Interest.na.z` mogen we in SD’s interpreteren aangezien ook deze variabele is gestandaardiseerd.) Aangezien de p-waarde kleiner is dan 0.05 verwachten we dit effect ook in de populatie terug te vinden. Hetzelfde geldt voor het effect van `Projectinteressant.z`. Eén SD hoger score op deze variabele leidt tot een toename van 0.188 SD op `Interest.na.z` ($= 0.188$, $p < 0.001$), ongeacht de studierichting waarin een leerling les volgt en zijn score op `Interest.voor.z`.

Ten slotte blijkt dat dit model 43.9% van de variantie in `Interest.na.z` verklaart (adj. R-kwadraat = 0.439). De F-toets leert dat we dit ook in de populatie mogen verwachten ($F(6,2236) = 293.2$, $p < 0.001$). Dus ook de in populatie verklaren de onafhankelijke variabelen ene deel van de verschillen in interesse in techniek na het project.

Oefening 4

Oefening 4

Ten slotte verwachten we dat het effect van `Interest.voor.z` op interesse in techniek mogelijk anders is voor leerlingen uit de studierichting “Technisch”.

- (a) Ga ook deze hypothese na. Rapporteer je bevindingen a.d.h.v. de output van je analyses.
- (b) Bereken de scores van volgende leerlingen in de steekproef en in de populatie:
 - een leerling die 0 scoort op alle onafhankelijke variabelen
 - een leerling uit een technische richting die 1SD hoger scoort op `Interest.voor.z` en 2 SD hoger scoort op `Projectinteressant.z`

(a)

Om deze onderzoeksvraag op een efficiënte wijze te vertalen in een model is het handiger om een andere referentiecategorie te nemen voor studierichting dan **Technisch**. In de code hieronder kiezen we ervoor om een andere grotere groep leerlingen te hanteren als referentiecategorie: de leerlingen uit Moderne Wetenschappen. Dit zie je in de code omdat de dummy variabele `ModWetniet` is opgenomen in het model. De vraag naar het interactie-effect vertalen we in het model door de term `(Interest.voor.z * Technisch)` op te nemen.

```
Model4 <- lm(
  Interest.na.z ~ Projectinteressant.z +
    Interest.voor.z +
    Technisch +
    Latijn +
    Kunst +
    Handel +
    (Interest.voor.z * Technisch),
  data=Techniek
)

summary(Model4)
```

Call:

```
lm(formula = Interest.na.z ~ Projectinteressant.z + Interest.voor.z +
    Technisch + Latijn + Kunst + Handel + (Interest.voor.z *
```



```
Technisch), data = Techniek)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.9124	-0.4541	0.0083	0.4787	2.7545

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.030234	0.023577	-1.282	0.19985
Projectinteressant.z	0.187985	0.016437	11.437	< 2e-16 ***
Interest.voor.z	0.563815	0.017519	32.183	< 2e-16 ***
Technisch	0.253368	0.089634	2.827	0.00475 **
Latijn	-0.029504	0.036677	-0.804	0.42124
Kunst	0.282396	0.120925	2.335	0.01962 *
Handel	0.008868	0.049540	0.179	0.85794
Interest.voor.z:Technisch	0.066976	0.073794	0.908	0.36418

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7497 on 2235 degrees of freedom

(124 observations deleted due to missingness)

Multiple R-squared: 0.4405, Adjusted R-squared: 0.4388

F-statistic: 251.4 on 7 and 2235 DF, p-value: < 2.2e-16

De vooropgestelde hypothese klopt niet! Uit de analyse blijkt dat het interactie-effect tussen `Interest.voor.z` en een technische studierichting volgen (`Technisch`) niet statistisch significant is ($\beta = 0.067$, $p = 0.364$). Het effect van `Interest.voor.z` op `Interest.na.z` is in de populatie dus niet anders voor leerlingen uit een technische studierichting.

Note

Als dit wel statistisch significant zou zijn, dan zou dat betekenen dat voor leerlingen uit een technische studierichting het effect van `Interest.voor.z` op `Interest.na.z` groter zou zijn. 1 SD hoger scoren op `Interest.voor.z` leidt dan voor een leerling uit een technische studierichting tot een toename van 0.631 ($= 0.564 + 0.067$) SD op `Interest.na.z`.

(b)

“Een leerling die 0 scoort op alle onafhankelijke variabelen”

... is een leerling die gemiddeld scoort op alle onafhankelijke kwantitatieve variabelen en die tot de referentiecategorie behoort (Moderne Wetenschappen). Onze beste inschatting is dat

deze leerling zowel in de steekproef een score op `Interest.na.z` van -0.03 (= het intercept!) behaalt. Voor de populatie verwachten we dat zo'n leerling een score 0 behaalt (het intercept is immers niet statistisch significant).

“Een leerling uit een technische richting die 1 SD hoger scoort op `Interest.voor.z` en 2 SD hoger scoort op `Projectinteressant.z`”

Vooraleer we de berekeningen doen, schrijven we de regressievergelijking opnieuw op.

$$\begin{aligned} \text{Interest.na.Z} = & \beta_0 + \beta_1 \cdot \text{Projectinteressant.z} + \beta_2 \cdot \text{Interest.voor.z} + \\ & \beta_3 \cdot \text{Technisch} + \beta_4 \cdot \text{Latijn} + \beta_5 \cdot \text{Kunst} + \beta_6 \cdot \text{Handel} + \\ & \beta_7 \cdot (\text{Technisch} * \text{Interest.voor.z}) \end{aligned} \quad (1)$$

Voorspelde scores **voor de steekproef**:

Om de scores te berekenen voor de steekproef vullen we alle parameters in uit de output.

$$\begin{aligned} \text{Interest.na.Z} = & -0.03 + 0.188 \cdot \text{Projectinteressant.z} + 0.564 \cdot \text{Interest.voor.z} + \\ & 0.253 \cdot \text{Technisch} + -0.030 \cdot \text{Latijn} + 0.282 \cdot \text{Kunst} + 0.009 \cdot \text{Handel} + \\ & 0.067 \cdot (\text{Technisch} * \text{Interest.voor.z}) \end{aligned} \quad (2)$$

Nu we de parameters hebben ingevuld, kunnen we ook de waarden voor de variabelen invullen die de betrokken leerling behaalt.

$$\begin{aligned} \text{Interest.na.Z} = & -0.03 + 0.188 \cdot 2 + 0.564 \cdot 1 + \\ & 0.253 \cdot 1 + -0.030 \cdot 0 + 0.282 \cdot 0 + 0.009 \cdot 0 + \\ & 0.067 \cdot (1 * 1) = 1.23 \end{aligned} \quad (3)$$

Voorspelde scores **voor de populatie**:

Om de scores te berekenen voor de populatie vullen we enkel de statistisch significante parameters uit de output in onze vergelijking in. Parameters die niet statistisch significant zijn vervangen we door de waarde 0.

$$\begin{aligned} \text{Interest.na.Z} = & 0 + 0.188 \cdot \text{Projectinteressant.z} + 0.564 \cdot \text{Interest.voor.z} + \\ & 0.253 \cdot \text{Technisch} + 0 \cdot \text{Latijn} + 0.282 \cdot \text{Kunst} + 0 \cdot \text{Handel} + \\ & 0 \cdot (\text{Technisch} * \text{Interest.voor.z}) \end{aligned} \quad (4)$$

Vervolgens kunnen we de waarden invullen die de bewuste leerling behaalde op elk van de variabelen in ons model.

$$\begin{aligned}
 \text{Interest.na.Z} = & 0 + 0.188 \cdot 2 + 0.564 \cdot 1 + \\
 & 0.253 \cdot 1 + 0 \cdot 0 + 0.282 \cdot 0 + 0 \cdot 0 + \\
 & 0 \cdot (1 * 1) = 1.193
 \end{aligned}
 \tag{5}$$