Statistiek 2 - Samenvatting

ASSUMPTIES

1. Independence: De observaties moeten onafhankelijk van elkaar zijn.
2. Interval scale: De dependent variabele moet gemeten zijn in een interval schaal
3. Normality: Elke steekproef is gevonden van een normaal verdeelde populatie. Dit kan aangetoond worden met met normaal kwantiel plotten(QQ-plot) en de Shapiro-Wilk test
4. Homogeneity of variance: De groepen hebben dezelfde variantie. Dit kan aangetoond worden met de Levene's en Hartley's test.
5. Randomness: De cases moeten afgeleid zijn van een willekeurige steekproef en de scores van de verschillende deelnemers moeten onafhankelijk van elkaar zijn.
6. Spherity: Van de aantal levels die er zijn kunnen paren worden gemaakt. Van alle paren worden de verschillen berekend, die allemaal dezelfde variantie zouden moeten hebben. Dit kan aangetoond worden met de Mauchly's test.
7. Linearity: Maak een residual plot. Dit is een scatterplot van de regressie residuals tegen de explanatory variabele. Dit geeft ons de mogelijkheid om te oordelen over de fit van de regressielijn. In deze curve moet je patroon van linearity vinden, als je deze vindt, dan is de assumptie met.
8. Geen multicollinearity: Er hoort geen perfecte correlatie te zijn tussen de predictors. Als 2 predictor variabelen sterk gecorreleerd zijn, betekent dat dat je de ene waarde kan voorspellen aan de hand van de lineairity van de ander met hoge betrouwbaarheid. Het heeft verder geen effect op de predictive power of de betroubaarheid, maar wel op de resultaten van de individuele predictors en welke predictors overbodig (redundant) zijn ten opzicht van anderen. Je kan dit controlleren door de correlatie te bepalen van alle variabelen met elkaar. De correlatie moet dan lager dan 0,9 zijn. Ook moet je een scatterplot maken, wat logisch is. Je kan ook de VIF scores bepalen, deze moeten lager zijn dan 5.
9. Homoscedasticity: De variabiliteit van de gegevens moet ongeveer gelijk zijn aan het bereik van de predicted waarden. Op elk niveau van de predictors moet de variantie van de residuals constant zijn. De residual plot moet een blob-vorm hebben. Het moet een horizontale lijn rond 0 zijn.
10. Normality of residuals: Deze assumptie is niet zo belangrijk. Maak een QQ-plot en doe de Shapiro-Wilk test van de residuals.
11. Absence of influential datapoints: Bereken de Cook's distance. Om de assumptie te metten, moet deze lager zijn dan 1. Je moet niet direct outliers weghalen, daar moet je een goede substantieve reden voor hebben. Als je een van deze punten weghaalt, verandert je regressielijn aanzienlijk.
12. Residuals zijn normaal verdeeld (toon je aan met QQ-plot en Shapiro-Wilk test) en niet autogecorreleerd. Oftewel de residual moeten onafhankelijk zijn. Kun je aantonen met de durbinWatson test, waarbij p GROTER moet zijn dan 0,05 en DW dicht bij de 2 moet liggen voor geen auto correlatie.

De Shapiro-Wilk test test met de nulhypothese of de groepen normaal verdeeld zijn. Als de p-waarde GROTER is dan 0,05 is de groep normaal verdeeld. Maar als er grote groepen zijn met kleine standaardafwijkingen is de assumptie makkelijker aan te nemen. Als er ook datapunten dezelfde waarde hebben, kan dit ook invloed hebben.

Levene's test: Ook hier kan een p-waarde gevonden worden. Als de p-waarde GROTER is dan 0,05 hebben de groepen dezelfde variantie.
→ Als normality is getest, en de groepen niet normaal verdeeld zijn, kan er een alternatieve test gedaan worden. Namelijk de Fligner-Killeen median test.

Hartley's test: Deze test berekent de N en de variantie. Hierna moet je zelf berekenen: grootste variantie - laagste variantie. De kritische waarde kan worden opgezocht bij: DF (k, n-1), waarin k staat voor het aantal groepen en n voor het aantal observaties in totaal. Als de gevonden waarde LAGER is dan de kritische waarde, hebben de groepen dezelfde variantie.
→ Als alle andere assumpties bij de ANOVA wel aangenomen zijn, maar de homogeneity niet, dan kan in R de oneway.test() (Welch's F) gebruikt worden.

Mauchly's test: De nulhypothese geeft aan dat de varianties van de verschillen gelijk zijn. Als p GROTER is dan 0,05, is de assumptie met en zijn de varianties gelijk. Als deze assumptie niet kan worden aagenomen, gebruiken we de Greenhouse-Geisser conservative F-test ipv die van de repeaturd measures test.

ncvTest: Met deze test kun je de homoscedasticity aantonen. Als de p-waarde HOGER is dan 0,05 is er sprake van homoscedasticity.

TEST	WANNEER
t-test	Hebben 2 groepen hetzelfde gemiddelde?
One-way ANOVA	Hebben 3 groepen hetzelfde gemiddelde?
Factorial ANOVA	Hebben 3 groepen en verschillende factoren hetzelfde gemiddelde?
Repeated measure ANOVA	Zijn er meerdere metingen gedaan van dezelfde variabelen, zoals dezelfde patiënt, maar verschillende tijden?
Correlatie	Is er een verband tussen 2 numerieke variabelen?
Simple linear regression	Verklaar de gevonden relatie tussen de numerieke explanatory en respons variabele
Multiple linear regression	Verklaar de gevonden relatie tussen meer dan 1 numerieke explanatory variabelen en de respons variabele.
Logistic regression	Verklaar de gevonden relatie tussen de categorische explanatory variabele(n) en de respons variabele.

Mixed-effect model	Wanneer je een ANOVA wil toepassen met ontbrekende waarden

	one-way ANOVA	factorial ANOVA	repeated ANOVA	simple regression	multiple regression	logistic regression	mixed-effect model
Independence	x	x		x	x	x	
Intervalschaal	x	x			x		
Normality	x	x	x		x		x
Homogeneity	x	x	x				
Randomness			x				
Sphericity			x				
Linearity				x	x	x	
No multicollinearity				x	x	x	
Homoscedasticity				x	x		x
Normality of residuals				x			x
Influential datapoints				x			
Autocorrelated					x		

One-way ANOVA

- Waarom doen we een ANOVA en geen t-test?

Het probleem is *inflation of suprise.* Dit houdt in dat de kans op een type 1 fout vergroot wordt. De kans op het vinden van een verrassend resultaat stijgt. Als er 3 groepen zijn, die onafhankelijk zijn, houdt dat in dat de kans op een type 1 fout: 0,95*0,95*0,95 = 0,857 is. De kans op een type 1 fout is dan: 1 - 0,857 = 0,143 * 100% = 14,3%, terwijl deze kans anders maar 5% was.
Type 1 error: H0 verwerpen, terwijl deze aangenomen moet worden.
Type 2 error: H0 accepteren, terwijl deze niet aangenomen moet worden.

Boxplotten laten de medianen zijn, niet de gemiddelden. Door te kijken naar de gemiddelden, standaardafwijkingen, de grootte van de steekproeven en de side-by-side boxplotten kunnen we kijken of de verschillen significant zijn.

<u>Hypotheses</u>
H0: $\mu 1 = \mu 2 = \mu 3$
Ha: not all of the μi are equal
Als de H0 hypothese is verworpen, kunnen we verder analyseren met contrasten en multi-comparisons.

Opmerkingen assumpties

- Bij een ANOVA test je dus of verschillende groepen verschillend zijn van 1 factor. Deze factor is de dependent variabele. Voor de assumpties moeten de groepen dus allemaal onafhankelijk van elkaar zijn.
- Bij de QQ-plotten, maak je de plot en door middel van QQ-line krijg je er een lijn doorheen. De punten moeten ongeveer lineair lopen met de QQ-line om normaal verdeeld te zijn.

Als we aannemen dat de populatie standaardafwijkingen gelijk zijn, dan zijn de steekproef standaardafwijkingen schatting van σ. We kunnen deze combineren tot 1 schatting: de gepoolde variantie σ^2, wat een onpartijdige (unbiased) schatting is.

We zoeken een algemeen patroon en afwijkingen ervan, waarbij geldt: DATA = FIT + RESIDUALS, wat in dit geval hetzelfde is als:
totale variantie = variantie between groepen + variantie within groepen (SST = SSG + SSE)

- De variantie between groepen is de SSG (Sum of Squares Group). De vrijheidsgraden zijn I - 1.
- De variantie within groepen is de SSE (Sum of Squares Error). De vrijheidsgraden zijn N - I.
- De variantie totaal is de SST (Sum of Squares Total). De vrijheidsgraden zijn N - 1.

Door de sum of squares te delen door de vrijheidsgraden, krijg je de mean squares (MSG, MSE en MST). De ANOVA berekent de F-toets, waarin MSG gedeeld wordt door MSE.

Als H0 waar is, heeft de F statistiek een F distributie die afhankelijk is van 2 getallen: de vrijheidsgraden van de numerator en die van de denominator: F (I - 1, N - 1). Als F KLEINER is dan de kritische waarde, wordt H0 aangenomen en zijn er geen verschillen tussen de gemiddelden.

- Als er een grote F-waarde gevonden wordt is Ha waar.

Als we Ha aannemen, kunnen we de data verder analyseren. We kunnen dan kijken welke groepen een verschillend gemiddelde hebben van elkaar. Als je dit met contrasten wil vergelijken, moet je van te voren hebben geformuleerd welke vragen je interessant vindt. Als je geen specifieke relaties in gedachten hebt, kun je gebruik maken van multi-comparisons.

Contrasts

Hypotheses

H0 : $\mu 1 = \frac{1}{2} (\mu 2 + \mu 3)$
Ha : $\mu 1 < \frac{1}{2} (\mu 2 + \mu 3)$
Deze hypothese kun je opnieuw formuleren tot:
$1\ \mu 1 - 0{,}5\ \mu 2 - 0{,}5\ \mu 3 = 0$
Dit kan berekend worden met een test in R. Geen idee welke. De t-test geeft een tweezijdige p-waarde. Als de alternatieve hypothese eenzijdig is, kun je de p-waarde delen door 2 om de juiste waarde te krijgen. Om H0 te accepteren, moet de p-waarde lager zijn dan 0,05.

Multi-comparisons

Voor deze test moeten we voor alle paren een t-test uitvoeren. Er kan een t** bepaald worden, welke afhankelijk is van welke procedure we kiezen. Als de nieuwe t kleiner is dan -t** of groter dan

t**, dan zijn de populatiegemiddelden verschillend. Anders kunnen we opmaken dat de data geen onderscheid maakt tussen beide groepen.

We kunnen kiezen voor de LSD (least-significant differences method). Deze procedure verlaagt de kans op een type I error. Hierbij is t** $\alpha/2$ critical value for the t(DFE) distribution.

Een andere optie is de Bonferroni methode. Hierbij is alpha verdeeld over het aantal vergelijkingen. Op deze manier is het alpha-level voor elk paar verlaagd, waardoor het alpha-level voor de hele analyse wel 0,05 blijft. Deze procedure verlaagt de kans op een type 2 error. De Holm-Bonferroni methode is een betere keuze, omdat deze meer power heeft.
* Dit kun je berekenen door te kijken naar het aantal groepen. Stel dat er 5 groepen zijn, dan kun je alle 5 de groepen met elkaar combineren en deze combinaties uitschrijven. Je ziet dan dat er 10 combinaties mogelijk zijn. Om dan hetzelfde alpha-level te houden voor de hele analyse, moet het oude alpha leven (0,05) door het aantal mogelijke combinaties gedeeld worden. Dus 10, waarmee het nieuwe alpha-leven 0,005 wordt.

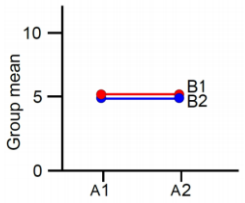
Betrouwbaarheidsintervallen geven aan dat we de verschillen niet precies weten. Simultaneously betrouwbaarheidsintervallen zijn intervallen voor alle verschillen tussen de populatiegemiddelden in 1. Deze hebben boven en ondergrenzen. De t** zijn hetzelfde. Als er geen significant verschil wordt gevonden tussen de gemiddelden, dan blijft het betrouwbaarheidsinterval 0.

Als met de Shapiro-Wilk test is aangetoond dat de data niet normaal verdeeld is, kan er een alternatieve test gebruikt worden. Dit is de Kruskal-Wallis test. In R: kruskal.test(). De hypotheses zijn:
H0: De populatie medianen zijn hetzelfde voor alle groepen
Ha: De populatie medianen zijn niet hetzelfde voor alle groepen

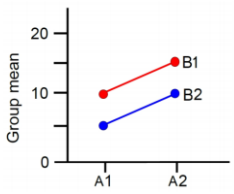
Factorial ANOVA

Voordelen aan de test:

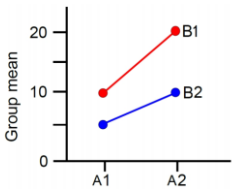
1. Het is efficiënter om meerdere factoren in 1x te bestuderen dan elke factor apart, omdat het aantal subjecten lager wordt. Je kan dan bijvoorbeeld 64 mensen testen ipv 80 in het voorbeeld.
2. De gepoolde variantie is lager bij een 2-factor ANOVA dan bij een 1-factor ANOVA. Oftewel een better prediction power. Dit komt omdat de tweede factor niet meer in de within groep zit, maar in de between groep. Elke keer kan de variantie verplaatsen van RESIDUAL naar FIT, waardoor de variantie kleiner wordt en de power van discriminatie toeneemt. Oftewel it is more likely to detect effects.
3. De interactie tussen factoren kan bekeken worden.



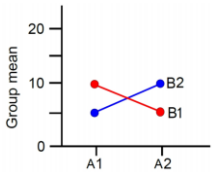
Geen main effect voor A of B
Geen interactie



Main effect voor A en B, omdat A2 hoger gemiddelde heeft dan A1
Geen interactie



Main effects voor A en B
Interactie, omdat B1 en B2 dichter bij elkaar liggen dan A1 en A2



Geen main effect, omdat ze elkaar neutraliseren.
Interactie

<u>Hypotheses</u>

Voor de hypotheses kan er gekozen worden uit meerdere opties. Het ligt eraan waar je in geïnteresseerd bent:

Main effect voor A:	*Main effect voor B*	*Interaction between A en B*
H0: μA1 = μA2 = μAI	H0: μB1 = μB2 = μBI	H0: μAB11 = μAB12 = μABij
Ha: niet alle μ zijn gelijk	Ha: niet alle μ zijn gelijk	Ha: niet alle μAB zijn gelijk

<u>Opmerkingen assumpties</u>

- Voor elke groep moet de normality bepaald worden. Dus als je 4 groepen hebt en 2 factoren, moet je dus voor 4x2= 8 groepen de Shapiro-Wilk test uitvoeren.
- Als eerste doe je de Levene's test, maar als geen significant resultaat oplevert, kan de Hartley's test gedaan worden. Hierin is de steekproefstandaardafwijking een benadering van σ. Deze willen we combineren tot 1 schatting. Hierbij kan de homogeneity toch gehaald worden.

Source	Sum of squares	Degrees of freedom	Mean sum of squares	F
Model	SSM	DFM=$(I \times J) - 1$	MSM=SSM/DFM	F=MSM/MSE
A	SSA	DFA=$I - 1$	MSA=SSA/DFA	F_A=MSA/MSE
B	SSB	DFB=$J - 1$	MSB=SSB/DFB	F_B=MSB/MSE
AB	SSAB	DFAB=$(I - 1) \times (J - 1)$	MSAB=SSAB/DFAB	F_{AB}=MSAB/MSE
Error	SSE	DFE=$N - (I \times J)$	MSE=SSE/DFE	
Totaal	SST	DFT=$N - 1$	MST=SST/DFT	

* Als er bijvoorbeeld 4 condities zijn en 2 factoren, staat elke factor voor 2 groepen/condities. Dit heeft invloed op de vrijheidsgraden dus hou hier rekening mee.

Als homogeneity niet behaald wordt, kan de White's adjustment worden toegevoegd. De code in R wordt dan: Anova(results.aov, type=''III'', white.adjust = TRUE)

Repeated measures

De observational units zijn datgene wat je meet. Er zijn verschillende metingen die je doet van hetzelfde. Je kan bijvoorbeeld iemand zijn gewicht iedere maand meten. Het lijkt op een gepaarde t-test met 3 of meer condities.

Het voordeel van repeated measures ANOVA tegenover onafhankelijke measure design (gewone ANOVA dus) is dat er minder subjects nodig zijn en dat deze methode more likely is to detect effects.

Het verschil bij repeated measures is de SSE. De variatie heeft verschillende oorzaken, zoals factoren waar de onderzoeker niks aan kan doen, zoals het weer enzovoort. Deze verschillen kunnen een grotere SSE geven voor individuele verschillen tussen de observational units. Daarom willen we deze weghalen uit de SSE en berekenen we een nieuwe SSE*. Deze is te berekenen volgens: SSE* = SSE - SSS, waarin SSS staat voor Sum of Squares Subject. De vrijheidsgraden zijn: DFE* = DFE - DFS. De F-toets wordt uiteindelijk: F = MSG/MSE*. Deze F zal groter zijn dan de F die we vonden in een one-way ANOVA.

Als Ha aangenomen wordt, kan daarna gekeken worden welke groepen er dan niet gelijk zijn. Dit wordt gedaan met de Bonferroni en Sidak modificaties. Bonferroni is meer conservatief dan Sidak. Tukey gebruiken we niet.

De niet-parametrische alternatieven zijn:

- Friedman test: Deze wordt gebruikt voor ordinale data. Als Ha aangenomen wordt, kan de Wilcoxon signed-rank test gebruikt worden met de Bonferroni correlatie.
- Cochran's Q test: Voor nominale data.

Correlatie

Een perfecte correlatie is 1 of -1. De parametrische correlatie test is Pearson en de non-parametrische is: Spearman of Kendall.
De dependent variabele of de respons variabele meet het resultaat van het onderzoek. De independent variabele of de explanatory- of predictor variabele probeert de geobserveerde resultaten te verklaren. De relatie tussen deze 2 kun je weergeven in een scatterplot. De relatie is sterk als de punten dicht langs de lijn liggen.

Pearson meet de richting en sterkte van een lineair verband tussen 2 kwantitatieve variabelen. Teken hierbij altijd een scatter plot. De r waarde is niet beinvloed door de kans in de unit van de 2 variabelen. Door middel van een scatterplot ontdekken of outliers en niet-lineaire relaties verband houden met een lage correlatie. Outliers hebben een groot effect op de correlatie.

De assumptie voor de test is dat de data bivariate normaal verdeeld is. Dat willen zeggen dat y normaal verdeeld is voor elke waarde van x, en x normaal verdeeld is voor elke waarde van y. Als de punten in de scatterplot een ellipsis vorm hebben, is dat een indicatie van een bivariate normale verdeling. Een groot effect van r geeft niet meteen een goede significantie aan.

<u>Hypotheses</u>
H0: r = 0 (geen lineair verband)
Ha: r > 0; p-waarde is P
Ha: r < 0; p waarde is P
Ha: r niet gelijk aan 0; p-waarde is 2P

Spurious correlations: Een sterke correlatie betekent niet meteen een causaal verband. Soms spelen verborgen variabelen een rol.

Alternatieve testen:
Spearman's rank correlation coefficient: Is hetzelfde als Pearson, maar dan tussen de gerankte variabelen.
Kendall: Voor kleinere datasets met smaller ties
Ook zijn er nog non-parametrische testen die je kan gebruiken als de data niet normaal verdeeld is. Deze zijn wel minder powerful.

Single lineaire regressie

Correlatie bepaalt de mate waarin x en y met elkaar een verband hebben. Als x en y worden omgedraaid, blijft r hetzelfde. Regressie beschrijft hoe de ene variabele met de andere verband houdt door middel van een formule. Hierbij verandert de formule wel als x en y worden omgedraaid. r wordt gebruikt om de slope van een formule te bepalen.

Een regressielijn is een rechte lijn welke de verandering van y beschrijft ten overstaande van x. In dit geval is y de responsvariabele en x de explanatory variabele: $y = ax + b$.

De verschillen tussen de predicted values en de observaties zijn de residuals. De eigenschappen van een regressielijn zijn:

- Een verandering van 1 standaardafwijking in x, komt overeen met een verandering van r standaardafwijkingen in y.
- De lijn gaat altijd door de gemiddelde x en y punten.

De least-squares zijn de de afstanden van de lijn tot aan het daadwerkelijke punt. Deze lijn wordt gemaakt door een schatting te maken van de echte regressielijn van de populatie.

De data is willekeurig gekozen uit een populatie en geeft een lineair verband. Als we het onderzoek opnieuw zouden uitvoeren, zouden we een andere steekproef krijgen, met een lineair verband dat iets anders is. Voor elke fixed x (een combinatie van p waarden voor elke xi 1 waarde), vinden we een respons y met een normale verdeling en een standaardafwijking.

Hypotheses

H0: B1 = 0
Ha: B1 > of < of niet gelijk aan 0
Als je > hebt, houdt het in dat er een positieve relatie is tussen de 2 variabelen.
Voor de t-test is het aantal vrijheidsgraden n-2.

Opmerkingen assumpties

- In simple-regressie hoef je de multicollinearity assumptie niet te checken, want die is er toch niet. Aangezien je maar 1 predictor hebt.
- Homoscedasticity en homogeneity is in principe op een dieper level hetzelfde. Ze betekenen allebei dat de variantie van de residuals overal hetzelfde is. Dat wil zeggen dat de residual variantie rond de predicted scores hetzelfde is voor alle predicted values.

- Outliers zorgen ervoor dat je R^2 lager wordt.

Vanuit de geobserveerde waarden uit een steekproef, kun je predictor values maken en die voorspellen de waarden van een populatie. Dit gebeurt dus bij regressie enzo.
Correlatie bekijkt of er een relatie is tussen de 2 variabelen en simple lineair regression verklaart het verband tussen de explanatory (x) en respons (y) variabele. Als er geen correlatie is, kun je ook geen simple lineair regression doen.

Multiple lineair regressie

De mean respons van een lineaire functie van de explanatory variabelen is:
$\mu y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta pxp$
Waarbij geldt dat voor elke fixed x, de respons y een normale verdeling en standaardafwijking heeft. Dit is de populatie gemiddelde response. De predictors zijn weergegeven met x1 en x2. De details van de regressielijn zijn gecompliceerder dan die van simple regression. Bij simple lineair regression wordt de least-squares line alleen gebruikt als basis voor een conclusie. De b-waarden bij multiple lineair regression zijn ook meer gecompliceerd.

De hypotheses zijn hetzelfde als bij simple lineair regression:
H0: B1 = 0
Ha: B1 > of < of niet gelijk aan 0
Als je > hebt, houdt het in dat er een positieve relatie is tussen de 2 variabelen.
Het aantal vrijheidsgraden is n-p-1.

<u>Opmerkingen assumpties</u>

- Linearity: De residual plot geeft de residuals aan met de regressiecoëfficiënt verandering (rode lijn) gebaseerd op de explanatory variabele. De datapunten is de groene lijn en de verandering tussen de lijnen is de afwijking van linearity.

Selectie van predictors:

- Overfitten: Er zijn teveel predictors, waardoor de outcome niet goed is. Dit werkt goed op een getraind steekproef, maar niet op een nieuwe steekproef. In andere woorden, het overfit de steekproef en het geldt niet voor de populatie. Dit gebeurt als er te weinig observaties zijn een teveel explanatory variabelen. Houd daarom aan dat je 10-15 observaties per explanatory variabele wil hebben. Als de optimism slope hoger is dan 0,05 is er sprake van overfitting.
- Underfitten: De belangrijke predictors zijn weggelaten.
- Alles in 1 regressie: Alle explanatory variabelen zijn tegelijk ingevoerd. Hierbij worden alle mogelijke combinaties geprobeerd en R zoekt de juiste.
- Forward selectie: De belangrijkste explanatory variabele wordt als eerste ingevoerd en daarna de andere een voor een. In R begin met je met een leeg model en voeg je alles daarna toe.
- Backward eleminatie: Je begint met alle explanatory variabelen en haalt ze 1 voor 1 weg, te begin met de minste belangrijke. In R begin met je met alle variabelen en haal je ze steeds weg.

De laatste 2 gebruiken heuristics, en garanderen niet meteen de beste combinatie. De beste keuze is all-in-one regression. De stepwise technieken worden beinvloed door de willekeurige variatie van de

data en zullen niet altijd hetzelfde resultaat geven bij herhaling. Met AIC (Akaike Information criterion) kijken we welk model beter is dan de andere. Hoe lager de AIC, hoe beter het model.

Onderdrukking: Er zijn 2 explanatory variabelen (x1 en x2) en een responsvariabele (y). X1 heeft een positieve correlatie met y, terwijl x2 een kleine tot geen correlatie heeft met y. Wanneer x2 in het model wordt meegenomen, vergroot de fit van het model en worden b1 en x1 hoger. Wanneer x2 niet is meegenomen in het model, wordt de relatie onderdrukt, omdat x1 en x2 in een bepaalde mate toch gerelateerd aan elkaar zijn.

ANOVA vs. lineair regression

De factoren en explanatory variabelen kun je met elkaar vergelijken. De factoren in ANOVA zijn categorisch, waarbij het aantal factoren oneindig is, zo kan leeftijd van 0 tot 100 gaan bijvoorbeeld. Elke waarde van een factor staat voor een groep, zo is bijvoorbeeld bij geslacht 1 man en 2 vrouw.

ANOVA: Als er een i aantal cases zijn en een xj aantal explanatory variabelen, staat iedere combinatie xij voor een groep. De waarden van yi en de responsvariabele y staan voor de observaties. De ŷi staat voor het groepsgemiddelde, y met streepje voor het globale gemiddelde van de yi observaties.

DATA = FIT + RESIDUAL
SST = SSM + SSE

- ANOVA: SSM meet de variatie van het groepsgemiddelde rond het globale gemiddelde
- Regression: SSM meet de variatie van de responsgemiddelde (ŷi) vergeleken met het globale gemiddelde (y met streepje).
- ANOVA: SSE meet de variatie van de individuele observaties vergeleken met de groepsgemiddelden
- Regressie: SSE meet de variatie van de individuele observaties (yi) vergelijken met hun gemiddelde (ŷi).
- ANOVA: SST meet de variatie van de individuele observaties vergeleken met de globale gemiddelden
- Regressie: SSE meet de variatie van de individuele observaties (yi) vergelijken met de globale gemiddelden (y met streepje).

De F-toets is dan MSM/MSE, waarbij H0 wordt aangenomen wanneer F groter of gelijk is aan de kritische waarde. De vrijheidsgraden zijn hierbij (p, n-p-1).

Hypotheses:
H0: $\beta 1 = \beta 2 = \beta p = 0$, waarbij geen enkele variabele een explanatory variabele of respons variabele is.
Ha: Er is tenminste 1 βj niet gelijk aan 0. Tenminste 1 van de explanatory variabele heeft een lineair verband met de responsvariabele.

Logistic regression

Er zijn 2 soort logistic regression:

- Binomial of dichotomous: 2 mogelijke outcomes
- Multinomial of polytomous: 3 of meer mogelijke outcomes

Logistic regression heeft dezelfde formule als lineaire regressie, maar hierbij is geen ŷ maar g(x):

g(x) = b0 + b1x1 + b2x2 + bpxp
g(x) is de logit of logodds van de outcome. Het reflecteert de verandering van de outcame A met de andere mogelijk outcome B, waarbij de dependent variabele 2 mogelijke waardes heeft.

DATA	r	schwa
upper	30	6
middle	20	74
lower	4	50

Het totaal is 184. Als je nu de probability van bijvoorbeeld de schwa wil berekenen, moet je voor alle klassen de schwa optellen en delen door het totaal, dus:
Probabilities of having [schwa] = 6 + 74 + 50 / 184 = 0,71
De odds kun je dan berekenen door de probabilities van beiden door elkaar te delen. Als je odds van [schwa] tot [r] wil berekenen krijg je dan: 0,71 / 0,29 = 2,45. Je kan dan concluderen dat de veranderingen in [schwa] 2,45 keer zo groot is als die van [r]. Als het aantal odds 1 is, hebben de variabelen dezelfde probability. Voor elke status kunnen de odds ook afzonderlijk worden bepaald. De odds van [schwa] voor de upper status is: 6 / (6 + 30) / 30 / (6 + 30) = 0,20.

De log odds zijn de logaritmische transformaties van de odds. Hiervoor gelden dezelfde regels, dus dat als de log odds 0 is, zijn de uitspraken in dit geval gelijk, wat betreft de probability. De input van de logaritmische functie kan een negatieve of positieve waarde zijn, maar de output is altijd tussen 0 en 1. Wanneer je hoge of lage waardes voor x hebt, kun je het beste de logaritmische functie gebruiken ipv de logaritmische regressie.

<u>Hypotheses</u>
H0: deviance of the model = deviance without predictors
Ha: deviance of the model is niet gelijk aan deviance without predictors
Als de p-waarde GROTER is dan 0,05, dan wordt H0 aangenomen.

De goodness of fit kan berekend worden. Dit wordt gedaan met Concordance index C: Het aantal keren dat het model een hogere probability voorspeld wanneer de outcome A is en het model A voorspeld. Hetzelfde geldt voor B. Dit kan bijvoorbeeld zo zijn, dat iemand die een bepaalde ziekte heeft gehad een grotere kans heeft om die ziekte weer te krijgen, dan iemand die de ziekte niet heeft gehad. De eisen zijn:

- C = 0,5 : Geen discriminatie
- C tussen de 0,7 en 0,8 : Acceptabel
- C tussen de 0,8 en 0,9 : Excellent
- C hoger dan 0,9 : Outstanding

Mixed effects regression

Fixed-effect factoren hebben een herhaalbare en klein aantal levels. Een random-effect factor heeft een niet-herhaalbare steekproef uit een grotere populatie. Deze test is handig wanneer er data

ontbreekt en je kan makkelijk testen of het nodig is om deze te zien als random-effect factoren. Er is ook geen gebalanceerd design nodig, zoals in repeated measures ANOVA.

Random-effect factoren zijn factoren die systematische variatie kunnen introduceren. Het is belangrijk om random slopes te testen. Random intercepts en slopes wil zeggen dat we iets kunnen toevoegen aan populatie intercept en slope. Dit zorgt ervoor dat we variantiestructuren in onze data kunnen aanbrengen. Parsimony is een single parameter model die alle random slopes en intercepts modelleert. Deze toevoegingen zijn BLUP (Best Lineair Unbiased Predictors). AIC bepaalt wanneer dit wel en niet nodig is, net zoals bij multiple regression. De intercepts en slopes wisselen nogal, wat invloed kan hebben op de overige waarden. lmer in R ontdekt zelf de random-effect structuren. BLUP en lmer vermijden overfitting en verbeteren de voorspelbaarheid door de toevoegingen.

In linear mixed-effect regressie volgen de fouten een normale verdeling met een gemiddelde van 0 en een standaardafwijking die in elke cel hetzelfde is en voor elke covariate. Als dit niet zo is, kun je 1000/Y of een logaritme van Y nemen.

Als de residuals niet normaal verdeeld zijn en niet homoscedastic, dan moet je de dependent variabele transformeren. Check ook op outliers.

Stepwise variabele selectie:

- Bevat random intercepts
- Voegt potentiële explanatory variabelen 1 voor 1 toe
- Niet-significante predictors worden gedropped.
- Kies alleen een complexer model als de AIC met minstens 2 verhoogd wordt.

Voor hypotheses is deze manier wat problematisch, omdat de p-waarden te laag zijn voor eventuele multiple comparisons ook. Dit kun je oplossen door naar AIC te kijken.
De t-waarde waar je mee vergelijkt is 2. Als de t-waarde lager is dan 2, zijn de waarden niet significant.

Effectgrootte:
R^2 tussen de 0,01 en 0,06 geeft een klein effect
R^2 tussen de 0,06 en 0,14 geeft een medium effect
R^2 groter dan 0,14 geeft een groot effect

R^2 kan groter worden door het toevoegen van niet-significante modellen. Daarom is er ook de *adjusted* R^2. Deze compenseert door te kijken naar het aantal explanatory (verklarende) termen in verhouding tot het aantal datapunten. Je kan deze waarde beter gebruiken. De adjusted R^2 is altijd gelijk of kleiner dan R^2 en kan ook negatief zijn. Dat kan ook met R^2 en houdt in dat de grafiek de andere kant op loopt.

R^2 meet alleen de effecgrootte in de steekproef. Dit is een partijdige (biased) schatting, wat inhoudt dat R^2 de variantie in de populatie overschat. De onpartijdige schatting kan bepaald worden met:
w^2 = SSG - (DFG x MSE) / MSE + SST
Hoe groter de steekproef, hoe kleiner de kans op partijdigheid.

EFFECTGROOTTE	Formule
one-way ANOVA	R^2 = SSG / SST
factorial ANOVA	R^2 = SSM / SST
factorial ANOVA, per factor en interactie	η^2 = SSA / SSA + SSE
repeated measure ANOVA	R^2 = SSG / SSG + SSE*
single linear regression	R^2 = s predictied y / s observed y
multiple linear regression	R^2 = SSM / SST