

# Onderzoeksmethoden II 2018-2019

## Zelfstudie: theorie wetenschappelijk onderzoek

**Onderzoeksvraag** = vraag → waarover gaat je onderzoek

**Hypothese** = stelling → verwachtingen over onderzoek

Belangrijk: eenduidige definities van de begrippen in je theorie!

→ wetenschappelijke aanpak

- Theorievorming (verzamelen van gegevens)
- Modelbouw (analyse van gegevens)
- Toetsing van resultaten (rapporteren)

### Verzamelen gegevens:

- Primaire gegevens heb je zelf verzameld
  - Secundaire gegevens je gebruikt bestaande gegevens  
→ alle gegevens van andere partijen dan de bron zelf
- Crossectie** onderzoek: 1 groep op 1 tijdstip  
**Longitudinaal** onderzoek: 1 groep over meerdere tijdstippen  
**Panneldata:** meerdere groepen over meerdere tijdstippen  
(bv elke maand, regelmatig)

### Bij primaire data:

- Interview
  - Vormen
    - Gestructureerde interviews (vragenlijst) → kwantitatief onderzoekinterview
    - Semi-gestructureerde “ (thema’s & vragen) → kwalitatief oi
    - Niet-gestructureerde “ = diepte-interview (improvisatie) → kwalitatief oi
  - Interactie-vorm
    - 1-op-1 basis: 1 persoon
    - 1-op-velen: groepsdynamiek  
→ groepsinterviews: 2 of meer pers  
→ **focusgroepen:** groepsinterview, onderwerp is duidelijk en nauwkeurig
- Vragenlijst
  - Vragenlijst opstellen
    - Lijstvragen: een of meerdere antwoorden uit lijst
    - Categorievragen: wederzijds uitsluitende antwoordcategorieën
    - Rangordevragen: relatievebelangrijkheid achterhalen
    - Kwantiteitsvragen: (zelf getal invullen) bv leeftijd
    - Rooster: antwoorden op 2 of meer gelijksoortige vragen
  - Schaal- of beoordelingsvragen: meningen verzamelen
  - 1) **Likertschaal:** je vraagt respondent in hoeverre ze het eens/oneens/neutraal zijn op een schaal van 4-5-6-7  
→ Een variabele waarbij de antwoorden worden aangeduid op een Likert schaal is een controle variabele
  - 2) **Numerieke beoordelingsschaal** (cijfer van 1 tot 10)
  - 3) **Semantische differentiaalschaal:** bolletje verslepen op een schaal

Interview of vragenlijst? → bepalende factoren	
INTERVIEW	VRAGENLIJST
Doel: verkennend onderzoek	Doel: beschrijvend of verklarend onderzoek
Beter voor persoonlijk contact	Beter als omvang van steekproef groot is
Beter als je veel tijd nodig hebt + veel vragen + open vragen	Minder kans op sociale wenselijkheid (sneller intieme informatie)

#### Hoe bepaal je steekproef:

- **Stochastisch** = representatief voor populatie
- **Niet-stochastisch** = niet representatief (gericht selecteren - moeilijk te vinden respondenten)

#### 4 fasen van stochastische steekproef:

1. **Steekproefkader:** complete lijst van alle cases in populatie
2. **Steekproefomvang:** compromis tussen nauwkeurigheid en beschikbare tijd  
(Hoe kleiner de steekproefomvang, hoe groter nauwkeurigheid)
3. **Steekproefmethode** kiezen
4. Controleren of steekproef **representatief** is

#### → Steekproefmethodes

- Stochastische SPmethodes
  - **Enkelvoudig aselect** random mensen uit België
  - **Systematisch** 1 case kiezen om de 10 bezoekers van de bib
  - **Gestratificeerd aselect** uit elke provincie aselect mensen kiezen
  - **Cluster** provincies → random provincies eruit kiezen
  - **Getrapt** provincies → gemeenten → enkele gemeenten
- Niet-stochastische SPmethodes
  - **Quota** vereisten vooropstellen (zoeken naar eig)
  - **Doelgericht** op eigen oordeel cases kiezen
  - **Sneeuwbal** contact met iemand, geeft je 5 nieuwe contacten
  - **Zelfselecterend** zoekertje op FB, mensen gaan er zelf op in
  - **Gemakssteekproef** cases die gemakkelijk voor SP te krijgen zijn

#### Betrouwbaarheid & validiteit

- **Validiteit** nauwkeurigheid (geldigheid)
  - geeft vaker problemen bij vragenlijsten (interpretatie vragen)
    - **Interne validiteit** mate waarin het redeneren binnen het onderzoek correct is uitgevoerd
    - **Externe validiteit** mate waarin de resultaten te generaliseren zijn en niet enkel opgaan voor de testgroep
- **Betrouwbaarheid** robuustheid: geen fouten (bias) + consistente resultaten
  - geeft vaker problemen bij interviews
    - **Interviewerbias:** beïnvloeding door gedrag, opmerkingen, toon interviewer
    - **Respondentebias:** sociaal wenselijk gedrag van respondent

Wetenschappelijke integriteit: zorgvuldigheid, voorzichtigheid, betrouwbaarheid, verifieerbaarheid, onpartijdigheid, onafhankelijkheid

# PPT: introductie

## Inhoudelijk

### Theoretisch

- ✓ Verzamelen en analyse van gegevens: zelfstudie
- ✓ Multivariate regressie

### Toepassingen

- ✓ Afname en invoer enquête
- ✓ Data manipulatie in SPSS
- ✓ Bivariate analyses in SPSS (bv verschiltoetsen, chi-kwadraat toets)
- ✓ Regressie-analyse in SPSS

## Waarom data manipuleren?

- Hercoderen van variabelen  
“Recode into different variables” → antwoordcategorieën samen nemen (dummy)  
**Dummy variabelen** creëren: je neemt verschillende antwoordcategorieën van een variabele samen en geeft die een nieuwe waarde (bv oneens ‘2’ & eerder oneens ‘3’ = dummy ‘0’)
- Berekenen van variabelen (gemiddeldes, verschillen berekenen)
- Selecteren van respondenten (mannen – vrouwen)  
1 geval selecteren → bv regionbe = 2  
2 of meer gevallen → bv ANY(regionbe,2,3)

## Beschrijving van gegevens

- ‘frequencies’ → nominale & ordinale variabelen
- ‘descriptives’ → schaal- & ordinale variabelen die je interpreteert als schaal

## Hoe rapporteren?

1. Formuleer hypothese
2. Welke variabelen gebruik je
3. Welke soort variabelen (beschrijving via frequencies of descriptives?)
4. Tabel (met titel en N = aantal gegevens)
5. Verklaring tabel: belangrijkste vaststellingen

Afhankelijke variabele: wat je wil onderzoeken, is veranderlijk (bv geluksgevoel)

Onafhankelijke variabele: wat gegeven is en je niet kan veranderen of kiezen (bv geslacht)

# PPT: Bivariate analyses

## Soorten variabelen

- Nominaal: geen specifieke volgorde
  - Dichotome variabele: maar 2 categorieën (geslacht)
- Ordinaal: specifieke volgorde (mening: eens, oneens)
- Schaal: vaste meeteenheid + volgorde
  - Interval: heeft geen nulpunt (temperatuur in °C)
  - Ratio: heeft een nulpunt (inkomen)



Categorische variabelen

## Onderzoek verbanden

1. Significant verband tussen variabelen?  
→ significantietoetsing: is het verband betekenisvol voor populatie?
2. Hoe sterk is het verband? → associatiematen
3. Wat is de aard/richting van het verband? → interpretatie resultaten

## Significantietoetsing

Nulhypothese  $H_0$ : er is geen significant verband in populatie

Alternatieve hypothese  $H_a$ : er is een significant verband in populatie

**P-waarde** = kans op extremere waarde dan je steekproefgrootte

→ hoe kleiner p, hoe extremer waarde en hoe signifikanter het verband

Significantiewaarde: indien niet vermeld is  $\alpha = 0,05$

p-waarde < significantiewaarde	p-waarde > significantiewaarde
Ho verwerpen	Ho aanvaarden
Verspreiding is significant	Verspreiding is niet significant
Variaties zijn niet gelijk	Variaties zijn gelijk

## 3 soorten toetsen

### 1. Chi-kwadraat toets

- 2 nominale variabelen
- 2 ordinale variabelen
- 1 ordinale & 1 nominale variabele

### 2. Correlatie

- Pearson correlatie
  - 2 schaal variabelen
- Spearman rangcorrelatie
  - 2 ordinale variabelen
  - 1 ordinale & 1 schaal variabele

Hoe toets kiezen?	
Type vraag	Methode kiezen
Afhankelijkheid	→ chi-kwadraat
Correlatie	→ Pearson/Spearman
Gemiddelde	→ verschiltoets

### Associatiematen + sterkte & richting

!!! Enkel als er een significant verband is

- **Cramer's V** → voor Chi-kwadraat analyse  
0 = geen samenhang      1 = perfecte samenhang  
toont enkel sterkte, voor richting moet je kijken naar kruistabel
- **Correlatiecoëfficiënt** → voor correlatie  
-1 = perfect negatief      0 = perfect onafhankelijk      1 = perfect positief  
toont sterkte én richting van verband  
→ positieve waarde = als ene var stijgt dan stijgt andere ook  
→ negatieve waarde = als ene var stijgt dan daalt andere

### 3. Toetsen van hypothesen & verschiltoetsen

Toetsen: beslissen of een vooraf geformuleerde uitspraak ( $H_0$ ) juist of onjuist is

- 1<sup>e</sup> onderscheid = soort steekproef

- **onafhankelijke steekproeven**: 1 variabele vergelijken over 2 groepen (M/V)
- **afhankelijke steekproeven**: 2 variabelen vergelijken in 1 groep (voor-na)  
= gepaarde waarnemingen

- 2<sup>e</sup> onderscheid = variabele (waarvoor we naar gemiddelde/gemiddelde rang kijken)

		Parametrische toets	Niet-parametrische toets
Gemiddelde	Schaalvariabele	Als $N > 30$	Als $N < 30$
	Ordinale variabele geïnterpreteerd als schaalvariabele		
Gemiddelde rang	Ordinale variabele	////////////////////////////////	Altijd! Rangordetoets

Bij verschiltoetsen: de variabele die de indeling maakt is **nominaal of ordinaal** (bv M/V)

→ Van zodra je vraag over gemiddelde of gemiddelde rang gaat, kies je voor verschiltoets!

Soorten testen: (namen niet kennen, enkel onderscheid en uitvoering)

	Parametrische test	Niet-parametrische test
1 variabele (vgl met hypothetische waarde)	One-sample t test	Wilcoxon test
2 <b>onafhankelijke</b> groepen	Unpaired (independent samples) t test	Mann – Whitney test
2 <b>afhankelijke</b> groepen → gepaarde waarnemingen!!	Paired t test	Wilcoxon test
3+ <b>onafhankelijke</b> groepen	One-way ANOVA (variantie-analyse)	Kruskal-Wallis test
3+ <b>afhankelijke</b> groepen	Repeated measures ANOVA	Friedman test

### Kenmerken van toetsen

1. enkel t-toetsen, geen z-toetsen

- **T-toets = verschiltoets**  
→ gebruikt om na te gaan of het (populatie-)gemiddelde van een normaal verdeelde grootte afwijkt van een bepaalde waarde & of er een verschil is tussen de gemiddelden van twee groepen in de populatie
- in de praktijk: standaardafwijking zelden gegeven, dus wordt die geschat door de steekproefstandaardafwijking
- z-toets: normale verdeling  
t-toets: t-verdeling



t-verdeling met grote  $N$  = z-verdeling

Twee-zijdige p-waarden

In SPSS krijg je altijd twee-zijdige p-waarden die je moet vergelijken met  $\alpha$

Als je een eenzijdige p-waarde nodig hebt moet je p-waarde delen door 2

Extremere waarde:  $P(\bar{x} > \text{steekproefgrootte}) \rightarrow p\text{-waarde}/2$

Hier omgekeerd:  $P(\bar{x} < \text{steekproefgrootte}) \rightarrow 1 - (p\text{-waarde}/2)$

## Interpretaties

### Chi-kwadraat toets

Ho = onafhankelijkheid tussen de 2 variabelen

- Ordinale variabele (mening)
- Nominale variabele (ja-nee)

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	105,603 <sup>a</sup>	4	,000
Likelihood Ratio	105,113	4	,000
Linear-by-Linear Association	102,772	1	,000
N of Valid Cases	1861		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 27,26.

Gays and lesbians free to live life as they wish * Belonging to particular religion or denomination Crosstabulation					
			Belonging to particular religion or denomination		Total
			Yes	No	
Gays and lesbians free to live life as they wish	Agree strongly	Count	300	666	966
		% within Belonging to particular religion or denomination	40,2%	59,7%	51,9%
	Agree	Count	273	354	627
		% within Belonging to particular religion or denomination	36,6%	31,7%	33,7%
	Neither agree nor disagree	Count	74	51	125
		% within Belonging to particular religion or denomination	9,9%	4,6%	6,7%
	Disagree	Count	51	24	75
		% within Belonging to particular religion or denomination	6,8%	2,2%	4,0%
	Disagree strongly	Count	48	20	68
		% within Belonging to particular religion or denomination	6,4%	1,8%	3,7%
Total	Count	746	1115	1861	
	% within Belonging to particular religion or denomination	100,0%	100,0%	100,0%	

### Pearson Chi-Square

- **Asymp. Sig. = p-waarde**

SPSS geeft uitvoer in tweezijdige p-waarde (2-sided), deze mag je vergelijken met  $\alpha$ . Bij eenzijdige toets moet je p-waarde zelf berekenen dus delen door 2

→ p-waarde = kans dat Ho waar is

p-waarde = 0,000 < 0,05

→ Ho verwerpen = significant verband

- **Df = aantal vrijheidsgraden**

Geeft aantal waarden weer die vrij mogen variëren, om statistische parameters te benaderen

Df = 4 → (2-1)\*(5-1)

2 mog bij nominale var & 5 mog bij ordinale var

- **(Linear-by-linear association)**

Doet bijna hetzelfde als Chi kwadraat, behalve dat er een lineaire regressie is met 1 covariaat (ordinale variabele). Een covariaat is een factor die naar verwachting invloed heeft op de relatie tussen de onafhankelijke en afhankelijke variabele en waarvoor je controleert in de analyse.

De covariaat is hier 'mening'. Er wordt vanuit gegaan dat de kans dat er versprongen wordt tussen 'agree strongly' en 'agree' dezelfde is als tussen 'agree' en 'neutral' etc. Daarom is er maar 1 vrijheidsgraad.

- **Cramer's V**  
**sterkte verband (0 – 1)**

Hier is Cramer's V = 0,238 → zwak verband tussen variabelen (neigt meer naar 0)

Voor richting kijken we naar de kruistabel: vooral gelovigen zijn het oneens met de stelling, de ongelovigen minder.

Symmetric Measures		Value	Approx. Sig.
Nominal by Nominal	Phi	,238	,000
	Cramer's V	,238	,000
N of Valid Cases		1861	

### Pearson correlatie

Ho = geen correlatie tussen de 2 variabelen

- Schaalvariabele (leeftijd)
- Ordinale variabele gezien als schaalvariabele: van 0-10 (tevredenheid)

Correlations			
		Age of respondent, calculated	How satisfied with job
Age of respondent, calculated	Pearson Correlation	1	,119**
	Sig. (2-tailed)		,000
	N	1869	952
How satisfied with job	Pearson Correlation	,119**	1
	Sig. (2-tailed)	,000	
	N	952	952

\*\* . Correlation is significant at the 0.01 level (2-tailed).

### Pearson correlation

- **Correlatiecoëfficiënt**  
 -1 = perfect negatief      0 = perfect onafh      1 = perfect positief  
 Correlatiecoëfficiënt = 0,119 → zwakke correlatie  
 Richting: positief → als men ouder wordt stijgt de jobtevredenheid en omgekeerd
- **Sig. (2-tailed) = P-waarde**  
 P-waarde = 0,000 < 0,01 → Ho verwerpen → significante correlatie

### Spearman Rangcorrelatie

= maat voor samenhang tussen 2 variabelen op basis van rangnummers

→ voor elk rangnummer neem je verschil tussen de waarden

- 1 = perfecte overeenstemming: alle rangnummers komen overeen (2x beste)
- -1 = perfecte onenigheid: hoge vs lage waarden voor alle rangnummers (1x beste, 1x slechtste)

### Spearman's rho

#### Voorbeeld 1

- Sig. (2 tailed) = p-waarde  
 → altijd eerst checken of er een significant verband is!  
 p-waarde = 0,002 < 0,01  
 → Ho verwerpen  
 → significante correlatie
- Correlation coëfficiënt = -0,099  
 → heel zwakke correlatie  
 → richting = negatief → Hoe groter onderneming, hoe minder tevreden (en omgekeerd)
- Vermeld de juiste N in je rapportering! Dit is de N waarbij je voor beide gekozen variabelen data hebt. Hier: N = 947

Correlations				
			How satisfied with job	Establishment size
Spearman's rho	How satisfied with job	Correlation Coefficient	1,000	-,099**
		Sig. (2-tailed)	.	,002
		N	952	947
	Establishment size	Correlation Coefficient	-,099**	1,000
		Sig. (2-tailed)	,002	.
		N	947	1668

\*\* . Correlation is significant at the 0.01 level (2-tailed).

#### Voorbeeld 2

10 nieuwe automodellen worden beoordeeld door 2 consumententijdschriften (ranking van 1 'beste' tot 10 'slechtste')

P-waarde = 0,000 < 0,05 → Ho (er is geen correlatie in beoordeling) verwerpen → significant verband

Correlation coefficient = 0,939 → dicht bij 1: heel sterke positieve correlatie → de 2 tijdschriften zijn het grotendeels met elkaar eens

Model	T1	T2	d (verschil)
1	4	5	-1
2	1	2	-1
3	9	10	-1
4	5	6	-1
5	2	1	1
6	10	9	1
7	7	7	0
8	3	3	0
9	6	4	2
10	8	8	0

## Parametrische toetsen

### Independent samples t-test (unpaired t test)

=> 2 onafhankelijke groepen

“Is er een verschil tussen het gemiddelde vertrouwen in het nationale parlement tussen de Vlaamse en Waalse regio?”

- 1 variabele (vertrouwen)  
→ ordinale variabele geïnterpreteerd als schaalvariabele
- 2 onafhankelijke groepen (Vlamingen + Walen)  
→ verdeling via nominale variabele
- $H_0$  = gemiddeldes (vertrouwen) van beide groepen zijn gelijk  
 $H_a$  = gemiddeldes van de groepen verschillen
- Tweezijdige toets (geen richting in de vraag)

#### Group statistics

N = aantal gegevens

Mean = gemiddelde

Std. Deviation = standaardafwijking

Std. Error Mean = standaardfout van het gemiddelde

Group Statistics					
Region, Belgium		N	Mean	Std. Deviation	Std. Error Mean
Trust in country's parliament	Flemish region	1101	5,02	2,107	,063
	Walloon region	590	4,84	2,384	,098

#### Independent Samples Test

We willen weten of de gemiddeldes

gelijk zijn, hiervoor zullen we de *T-test for Equality of Means* gebruiken. Maar deze geeft 2 waarden, naargelang de assumptie of de varianties gelijk zijn of niet.

We onderzoeken dus eerst de varianties via de *Test for Equality of Variances*

		Levene's Test for Equality of Variances		t-test for Equality of Means				
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
Trust in country's parliament	Equal variances assumed	10,497	,001	1,614	1689	,107	,182	,113
	Equal variances not assumed			1,555	1083,4	,120	,182	,117

#### 1. Test for Equality of Variances

= testen of 2 normaal verdeelde populaties dezelfde variantie hebben

- $H_0$  = de varianties zijn gelijk  
 $H_a$  = de varianties zijn niet gelijk
- interpretatie van F  
Als  $F=1$  dan zijn de varianties perfect gelijk  
Hoe verder F is van 1, hoe meer waarschijnlijk dat de varianties verschillen  
 $F = 10,497 \rightarrow$  grote waarschijnlijkheid dat varianties verschillen
- Sig. = p-waarde

p-waarde > $\alpha$	p-waarde < $\alpha$
→ $H_0$ aanvaarden	→ $H_0$ verwerpen
→ varianties zijn gelijk	→ varianties zijn niet gelijk

$P\text{-waarde} = 0,001 < 0,05 \rightarrow H_0$  verwerpen → varianties zijn niet gelijk

→ We zitten dus in de situatie: 'Equal variances not assumed'



## 2. T-Test for Equality of Means

= testen of de 2 normaal verdeelde populaties dezelfde gemiddeldes hebben

- $H_0$  = de gemiddeldes zijn gelijk  
 $H_a$  = de gemiddeldes zijn niet gelijk
- Interpretatie van T  
Als  $T=0$ , dan zijn de gemiddeldes gelijk  
Hoe verder de waarde van T van 0, hoe meer waarschijnlijk dat er een significant verschil is  
 $T = 1,555 \rightarrow$  kleine kans dat gemiddeldes verschillen
- Df = aantal vrijheidsgraden
- Sig. (2 tailed) = tweezijdige p-waarde  
 $P\text{-waarde} = 0,120 > 0,05 \rightarrow H_0$  aanvaarden  
 $\rightarrow$  er is geen significant verschil tussen de gemiddeldes
- Mean difference = verschil tussen de gemiddeldes (niet significant hier)
- Std. Error Difference = verschil tussen de standaardfouten

## Niet-parametrische toetsen

### De Mann-Whitney toets (Rangsommentoets van Wilcoxon)

=> 2 onafhankelijke groepen

"Is er een verschil in kansverdeling/ gemiddeldes tussen de voorspellingen voor beide groepen?"

- 1 variabele ( voorspellingen kosten levensonderhoud) → schaalvariabele
- 2 onafhankelijke groepen (overheidseconomen & uniefeeconomen) → verdeling via nominale variabele
- Ho = de kansverdelingen/gemiddeldes van de 2 populaties zijn gelijk  
Ha = de kansverdelingen/gemiddeldes van de 2 populaties verschillen

**Voorbeeld:** voorspelling kosten voor levensonderhoud

Overheidseconomen		Universiteitseconomen	
voorspelling	rangnummer	voorspelling	rangnummer
3,1	4	4,4	6
4,8	7	5,8	9
2,3	2	3,9	5
5,6	8	8,7	11
0,0	1	6,3	10
2,9	3	10,5	12
		10,8	13

- Je hebt een aantal **voorspellingen** voor elke groep
- Je geeft elke steekproefwaarneming een **rangnummer** (alsof ze allemaal uit dezelfde populatie komen), 1 = kleinst
- **Sum of ranks:** je maakt de **som** van de rangnummers voor beide groepen

T1 (overheid) = 25 & T2 (unief) = 66

- **Mean rank = gemiddelde rang**

Unief = 9,43 > Overheid = 4,17

→ Uniefeeconomen zouden meer kosten hebben voor levensonderhoud

Ranks				
	type_econoom	N	Mean Rank	Sum of Ranks
voorspelling	Uniefeeconomen	7	9,43	66,00
	Overheidseconomen	6	4,17	25,00
	Total	13		

- Toetsingsgrootheid** is gebaseerd op de totalen van de rangnummersommen voor beide steekproeven  
→ 2 rangnummersommen zijn vrijwel gelijk = kansverdelingen zijn identiek  
→ 2 rangnummersommen zijn heel verschillend = kansverdelingen zijn verschillend  
Verschil tussen 25 en 66 is vrij groot → **verschillende kansverdelingen**

- Toetsen of het **verschil groot genoeg** is

- interpretatie Mann-Whitney **U waarde**  
Kleinst mogelijke waarde van U is 0  
Als U=0 dan zullen alle waardes in de ene groep groter zijn dan de waarden uit de andere groep
- interpretatie van de Wilcoxon **W waarde**  
dit is de kleinste rangnummersom van de groepen
- Z = z-score** → de p-waarde is gebaseerd op de standaardnormale verdeling
- Asymp. Sig. (2-tailed)** = tweezijdige p-waarde  
P-waarde = 0,015 < 0,05 → Ho verwerpen  
→ de kansverdelingen van de groepen verschillen  
Conclusie: het verschil is groot genoeg!

Test Statistics <sup>a</sup>	
	voorspelling
Mann-Whitney U	4,000
Wilcoxon W	25,000
Z	-2,429
Asymp. Sig. (2-tailed)	,015
Exact Sig. [2*(1-tailed Sig.)]	,014 <sup>b</sup>

a. Grouping Variable:  
type\_econoom  
b. Not corrected for ties.

Two-tailed of one-tailed sig. ?

Two-tailed: als je praat over een verschil

One-tailed: als je onderzoekt of het 'groter dan' of 'kleiner dan' iets is

**Voorbeeld:** beoordeling van zachtheid van twee soorten papier door 10 consumenten

Beoordelaar	Produkt		Vershil	Absolute waarde	Rangnummer
	A	B	(A-B)	verschil	Abs w verschil
1	12	8	4	4	4,5
2	16	10	6	6	7
3	8	9	-1	1	1
4	10	8	2	2	2
5	19	12	7	7	8
6	14	17	-3	3	3
7	12	4	8	8	9
8	10	6	4	4	4,5
9	12	17	-5	5	6
10	16	4	12	12	10

### Rangtekentoets van Wilcoxon

=> 2 afhankelijke groepen = gepaarde waarnemingen

“Is er een verschil in de kansverdeling van de beoordelingen van beide producten?”

- 1 variabele (beoordeling) → schaalvariabele
- 2 groepen (product A en B) → nominale variabele
- $H_0$  = kansverdelingen van de beoordelingen van product A en B zijn hetzelfde  
 $H_a$  = kansverdelingen van de beoordelingen van product A en B verschillen

1. Je krijgt beoordelingen voor 2 groepen (producten), het verschil tussen die beoordelingen en de absolute waarde van het verschil (alles positief)
2. Je geeft rangnummers aan de absolute waardes van de verschillen (startend van 1 = kleinst) → als een getal 2x voorkomt dan geef je ze het gemiddelde van de 2 rangnummers  
2x 4 → Elk rangnummer 4,5
3. Je maakt de **som** van de rangnummers van de negatieve verschillen en de som van de rangnummers van de positieve verschillen  
 $T+ = 45$  en  $T- = 10$  →  $T+$  is hoger → Product A beter beoordeeld

Negatieve verschillen	Rangnr
-1	1
-3	3
-5	6
Som rangnummers $T- = 10$	
Positieve verschillen	Rangnr
4	4,5
6	7
2	2
7	8
8	9
4	4,5
12	10
Som rangnummers $T+ = 45$	

4. **Toetsingsgrootheid T** = kleinste van  $T+$  en  $T-$   
→ hoe kleiner T, hoe waarschijnlijker dat de 2 kansverdelingen verschillen  
 $T = 10$  → niet zo'n grote kans dat ze verschillen

		Ranks		
		N	Mean Rank	Sum of Ranks
product_A - product_B	Negative Ranks	3 <sup>a</sup>	3,33	10,00
	Positive Ranks	7 <sup>b</sup>	6,43	45,00
	Ties	0 <sup>c</sup>		
	Total	10		

a. product\_A < product\_B  
b. product\_A > product\_B  
c. product\_A = product\_B

5. Toetsen of het **verschil groot genoeg** is
  - Interpretatie N: zegt hoeveel negatieve, positieve, en dezelfde (ties) waarden er zijn
  - *Sum of ranks* = sommen van de positieve & negatieve rangnummers
  - Z = z-score
  - Asymp. Sig (2-tailed) = tweezijdige p-waarde  
P-waarde = 0,074 > 0,05  
→  $H_0$  aanvaarden  
→ kansverdelingen van beide beoordelingen zijn gelijk  
Conclusie: het verschil is niet groot genoeg!

Test Statistics <sup>a</sup>	
	product_A - product_B
Z	-1,785 <sup>b</sup>
Asymp. Sig. (2-tailed)	,074
a. Wilcoxon Signed Ranks Test	
b. Based on negative ranks.	

## Kruskal-Wallis H toets

=> 3 of meer onafhankelijke groepen

"Is er een verschil in de kansverdeling/gemiddeldes van de beschikbare bedden in de 3 ziekenhuizen?"

- 1 variabele (# lege bedden) → schaalvariabele
  - 3 groepen (ziekenhuizen) → nominale variabele
  - Ho = kansverdelingen van het aantal lege bedden zijn hetzelfde voor de 3 ziekenhuizen
- Ha = Er verschilt minstens 1 kansverdeling van de andere**

**Voorbeeld:** beschikbare bedden in 3 ziekenhuizen

Ziekenhuis 1		Ziekenhuis 2		Ziekenhuis 3	
bedden	rangnummer	bedden	rangnummer	bedden	rangnummer
6	5	34	25	13	9,5
38	27	28	19	35	26
3	2	42	30	19	15
17	13	13	9,5	4	3
11	8	40	29	29	20
30	21	31	22	0	1
15	11	9	7	7	6
16	12	32	23	33	24
25	17	39	28	18	14
5	4	27	18	24	16
R <sub>1</sub> =120		R <sub>2</sub> =210,5		R <sub>3</sub> =134,5	

- Je geeft opnieuw rangnummers aan alle waarden, ongeacht de groep en maakt de som van de rangnummers voor elke groep.
- Toetsingsgrootheid T** is gebaseerd op de rangnummersommen voor elke van de steekproeven  
→ Als de sommen van de rangnummers dicht bij elkaar liggen dan wijst dit erop dat de kansverdelingen gelijk zijn  
120, 134,5 en 210,5 → slechts 1 rangnummersom verschilt veel van de andere

- Toetsen of het **verschil groot genoeg** is

- interpretatie Kruskal-Wallis H waarde  
hoe groter de H waarde, hoe groter het verschil tussen de rangnummersommen
- Df = vrijheidsgraden  
Df = 2 (3-1)
- Asymp. Sig. = tweezijdige p-waarde  
P-waarde = 0,047 < 0,05 → Ho verwerpen  
→ Ha: er is minstens 1 kansverdeling van lege bedden die verschilt van de andere  
Conclusie: het verschil is groot genoeg!

### Ranks

ziekenhuis		N	Mean Rank
bedden	ziekenhuis 1	10	12,00
	ziekenhuis 2	10	21,05
	ziekenhuis 3	10	13,45
	Total	30	

### Test Statistics<sup>a,b</sup>

bedden	
Kruskal-Wallis H	6,099
df	2
Asymp. Sig.	,047

a. Kruskal Wallis Test

b. Grouping Variable:  
ziekenhuis

## OPMERKINGEN

Feeling of safety of walking alone in local area after dark \* Region, Belgium Crosstabulation

			Region, Belgium			Total
			Flemish region	Brussels region	Walloon region	
Feeling of safety of walking alone in local area after dark	Very safe	Count	219	36	130	385
		% within Region, Belgium	19,9%	21,7%	21,7%	20,6%
	Safe	Count	684	103	343	1130
		% within Region, Belgium	62,1%	62,0%	57,4%	60,6%
	Unsafe	Count	171	26	110	307
		% within Region, Belgium	15,5%	15,7%	18,4%	16,5%
	Very unsafe	Count	28	1	15	44
		% within Region, Belgium	2,5%	0,6%	2,5%	2,4%
Total	Count	1102	166	598	1866	
	% within Region, Belgium	100,0%	100,0%	100,0%	100,0%	

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	6,643 <sup>a</sup>	6	,355
Likelihood Ratio	7,540	6	,274
Linear-by-Linear Association	,030	1	,862
N of Valid Cases	1866		

a. 1 cells (8,3%) have expected count less than 5. The minimum expected count is 3,91.

Vb: Chi-kwadraat

Als er **cellen zijn met een te lage theoretische frequentie** moet je cellen samennemen!! (zie voetnota)

Je steekt de categorie met het laagste totaal (44, totaal van very unsafe) bij een aangrenzende (unsafe).

→ recode into different variables

+ Zie oefn in ppt: goed weten welke variabelen, welke van de 3 toetsen en welke specifieke toets

# Theorie: Meervoudige lineaire regressie (H13 handboek statistiek)

## 1. Meervoudig lineair regressiemodel

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + e$$

- $y$ : de te verklaren (afhankelijke) veranderlijke. Moet ALTIJD een schaalvariabele zijn.
- $x_1, x_2, \dots, x_k$ : de verklarende (onafhankelijke) veranderlijken  $\rightarrow$  invloed op  $y$ ?
- $E(y) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k$ : deterministisch deel
- $\beta_i$ : de bijdrage van de verklarende veranderlijke  $x_i$
- $\varepsilon$ : toevallige afwijking

### Onderscheid x-variabelen

- Kernvariabele: determinant die centraal is in je onderzoek
- Controlevariabelen: andere x variabelen.  
!!! x-variabelen mogen geen functie zijn van elkaar vb leeftijd en anciënniteit

### Analyseren van een meervoudig regressiemodel:

- Stap 1: bepaal de deterministische component  $\rightarrow$  welke x-variabelen kies je?
- Stap 2: schat de parameters  $\beta_i \rightarrow$  adhv een steekproef
- Stap 3: specificeer de kansverdeling van  $\varepsilon \rightarrow$  schatting maken van standaardafwijking
- Stap 4: controleer de aannames (veronderstellingen) rond  $\varepsilon$   
De toevallige afwijking  $\varepsilon$  heeft een kansverdeling met de volgende eigenschappen:
  1. De verwachting is gelijk aan 0.
  2. De variantie is gelijk aan  $\sigma^2$ .
  3. De kansverdeling is normaal
  4. Verschillende toevallige afwijkingen zijn onafhankelijk van elkaar.
- Stap 5: beoordeel de bruikbaarheid van het model
- Stap 6: gebruik het model

## 2. Het eerste-orde model

$$E(y) = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5$$

Eerste orde model: y-variabele verklaren door enkele x-variabelen (hier 5)

- $x_i$ 's zijn **kwantitatieve** veranderlijken die **geen** functie zijn van andere verklarende veranderlijken
- $\beta_i$  stelt de helling van de lijn voor die het verband weergeeft tussen  $y$  en  $x_i$  (andere  $x$ 'en = constant)
- $\beta_0$  is het snijpunt met de y-as (startwaarde)

### $\beta_i$ 's schatten

Volgens de **kleinste kwadraten methode**

Methode om bij een gegeven verzameling punten in het xy-vlak, die verondersteld worden ong. op een rechte lijn te liggen, de "best passende" lijn te bepalen. Best passen = totaal van de gekwadrateerde afwijkingen in verticale zin van de punten t.o.v. de lijn is zo klein mogelijk

$\rightarrow$  Kleinste-kwadratenvoorspellingsvergelijking

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \dots + \hat{\beta}_kx_k$$

$\rightarrow$  **SSE** = som van de kwadraten van de afwijking

Verschil tussen effectieve waarde en de voorspelling ervan

$\rightarrow$  willen we minimaliseren om schatting zo dicht mogelijk bij werkelijke waarde te krijgen

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

### βi's berekenen via SPSS

1. Spreidingsdiagrammen (scatter plots)
2. Regressiecoëfficiënten schatten
3. SSE bepalen via ANOVA
4. Standaardafwijking bepalen

#### Voorbeeld 1:

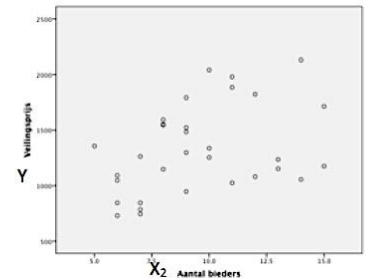
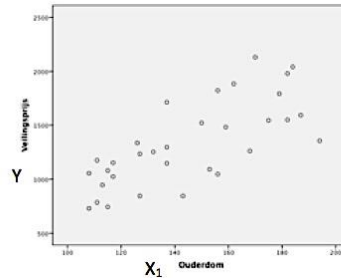
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

met:

- $y$  = veilingsprijs in euro's
- $x_1$  = ouderdom van de klok in jaren
- $x_2$  = aantal bidders

steekproefgrootte:  $n = 32$

1. **Scatter plot** tussen veilingsprijs  $y$ , ouderdom  $x_1$  en aantal bidders  $x_2$   
 → in 1<sup>e</sup> grafiek liggen waarnemingen meer op 1 lijn  
 → verband tussen verkoopprijs  $y$  en ouderdom  $x_1$  is sterker



2. Regressiecoëfficiënten

- **Unstandardized Coefficients: B**  
= **geschatte regressiecoëfficiënten**

- Standardized coefficients: Beta  
= tonen relatieve kracht van de regressiecoëfficiënten

- t-waarde:  
geeft betrouwbaarheid van de schatting van één individuele coëfficiënt weer  
→ dat de betreffende verklarende variabele iets toevoegt aan de verklaring van de variatie van  $y$   
→ significantie hiervoor bekijken

- Sig. = p-waarde → als p-waarde < 0,05 dan is je B coëfficiënt significant

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	β <sub>0</sub> (Constant)	-1338.951	173.809		-7.704	.000
	β <sub>1</sub> Ouderdom	12.741	.905	.887	14.082	.000
	β <sub>2</sub> Aantal bidders	85.953	8.729	.620	9.847	.000

a. Dependent Variable: Veilingsprijs

Kwadratenvoorspellingsvergelijking invullen

$$\hat{y} = -1338,95 + 12,74 * \text{ouderdom} + 85,95 * \text{bidders}$$

3. SSE bepalen via ANOVA

→ we willen SSE minimaliseren

- **Regression**

- Sum of Squares: totale variantie van alle waarnemingen

→ Mean Square: Sum of Squares / df

- F-waarde: hoge waarde = hoge variabiliteit van de gemiddeldes (gemiddeldes liggen ver van elkaar)

Sig. = 0,000 < 0,05 → het regressiemodel is significant (kan  $y$  goed voorspellen)

- **Residual Sum of Squares (SSE)**

= de minimumwaarde van de SSE = 516726,54

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4283062.96	2	2141531.48	120.188	.000 <sup>b</sup>
	Residual	516726.540	29	17818.157		
	Total	4799789.50	31			

a. Dependent Variable: Veilingsprijs

b. Predictors: (Constant), Aantal bidders, Ouderdom

#### 4. Standaardafwijking bepalen

**Residual Mean Square** = schatter van de afwijking ( $s^2$ )

$$s^2 = 17818,157 (= SSE/29) \rightarrow \text{zie output}$$

$$s = \sqrt{17818,157} = 133,5 \text{ (normaal ook in output)}$$

Interpretatie: het interval  $y \pm 2s$  is een grove schatting van de nauwkeurigheid waarmee men het model  $y$  voorspelt

$\rightarrow$  Hier zal men de veilingprijzen met ruwweg  $\pm 267$  eur nauwkeurigheid voorspellen.

Een schatter voor  $\sigma^2$  voor het meervoudig regressiemodel met  $k$  verklarende veranderlijken:

$$s^2 = \frac{SSE}{n - \text{aantal geschatte } \beta\text{'s}} = \frac{SSE}{n - (k+1)}$$

Voorbeeld:  $s^2 = 17818$

$$s = 133,5$$

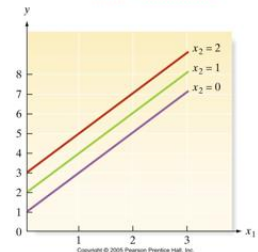
+ interpretatie

#### Wat is de impact op het model?

- Enkelvoudig regressie model  
Met hoeveel zal  $y$  toenemen als  $x$  toeneemt met 1 eenheid?
- Meervoudig regressie model  
Wat is het effect van  $x_1$  op  $y$  als  $x_2$  constant gehouden wordt?  
 $\rightarrow$  Effect van  $x_1$  op  $y$  blijft hetzelfde, ongeacht de waarde van  $x_2 \rightarrow$  evenwijdige lijnen

Gegeven het model  $E(y) = 1 + 2x_1 + x_2$ .

Wat is het effect van  $x_1$  op  $E(y)$ , wanneer  $x_2$  constant wordt gehouden?



#### Interpretatie van de geschatte $\beta$ 's

$\hat{\beta}_1 = 12,74$	$E(y)$ , de <u>verwachte verkoopprijs zal naar schatting toenemen met 12,74€ als de leeftijd van de klok met één jaar toeneemt (wanneer alle andere variabelen constant blijven).</u>
$\hat{\beta}_2 = 85,95$	$E(y)$ , de <u>verwachte verkoopprijs zal naar schatting toenemen met 85,95€ als het aantal bidders met één toeneemt (wanneer alle andere variabelen constant blijven).</u>

$\beta_0 \rightarrow$  vaak geen praktische betekenis

#### Een $100(1-\alpha)\%$ BI voor $\beta_i$ :

$$BI_{100(1-\alpha)\%} = \left[ \hat{\beta}_i \pm t_{n-(k+1)}^{\alpha/2} \cdot s_{\hat{\beta}_i} \right]$$

In SPSS output of zelf berekenen (hier):

$$BI_{90\%} = \left[ \hat{\beta}_1 \pm t_{32-(2+1)}^{0,05} \cdot s_{\hat{\beta}_1} \right]$$

$$\rightarrow \text{De verwachte veilingprijs neemt toe} \rightarrow \beta_2 > 0 \rightarrow [12,74 \pm 1,699 \cdot 0,905] = [11,21 ; 14,27]$$

#### Twee types gevolgtrekkingen voor $\beta_i$ :

- betrouwbaarheidsintervallen (BI)  
BI(90%) voor  $\beta_1$   
Ga na of  $\beta_2 > 0$  met  $\alpha=0.05$   
 $\rightarrow$  De verwachte veilingprijs neemt toe  $\rightarrow \beta_2 > 0$
- hypothesetoetsen  
Via SPSS moet men enkel de P-waarde interpreteren.  
 $P = 0,000 \rightarrow H_0$  verwerpen  
 $\rightarrow$  De verwachte veilingprijs neemt dus toe



### 3. Toetsen van de bruikbaarheid van een model

- Hoe goed past model bij gegevens? Hoe goed is schatting?:
  - meervoudige determinatiecoëfficiënt  $R^2$

$$R^2 = 1 - \frac{SSE}{SS_{yy}} = \frac{SS_{yy} - SSE}{SS_{yy}} = \frac{\text{Explained variability}}{\text{Total variability}}$$

- enkel gebruiken wanneer n aanzienlijk groter is dan (k+1)
- je moet meer cases hebben dan x-variabelen

$R^2$  toont de fractie van de variatie in y-waarde aan die door de regressielijn wordt verklaard.

- $SS_{yy}$  = in welke mate ga je afwijken van het gemiddelde (variatie in y) (je kijkt naar geobserveerde waarde en vergelijkt met gemiddelde)
  - $SSE$  = som van de geobserveerde y-waarde ten opzichte van de schatting (van y)
- Een lage  $R^2$  = X draagt weinig bij tot y ( $SS_{yy}$  en  $SSE$  liggen dicht bij elkaar)

- gecorrigeerde meervoudige determinatie-coëfficiënt  $R^2_a$

$$R^2_a = 1 - \left[ \frac{n-1}{n-k+1} \right] \left( \frac{SSE}{SS_{yy}} \right) = 1 - \left[ \frac{n-1}{n-k+1} \right] (1 - R^2)$$

- vergelijken over modellen heen
- gecorrigeerd voor aantal parameters en steekproefgrootte

- Bruikbaarheid van model → **Globale F-toets**
- als het niet bruikbaar is kan je niet verder gaan

$$F = \frac{s_x^2}{s_y^2} = \frac{\text{variantie van x}}{\text{variantie van y}}$$

**De globale F-toets (ANOVA) ≠ one way ANOVA**

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$  → *geen enkele waarde brengt iets bij tot het model*

$H_a$ : minstens één  $\beta_i \neq 0$  → hier is het bruikbaar

We kijken opnieuw naar de p-waarde. (als p-waarde <  $\alpha$  →  $H_a$  → model bruikbaar)

Globale bruikbaarheid van een meervoudig regressiemodel nagaan:

- **Stap 1:** voer de globale F-toets uit; wanneer  $H_0$  verworpen wordt, ga verder naar stap 2 (dit gaat louter bruikbaarheid na, niet of een model het 'beste' is!)
- **Stap 2:** voer t-toetsen uit op die  $\beta_i$ 's waarin je het meest geïnteresseerd bent

Bij F altijd grootste getal in de teller!

Als F = 1 dan zijn zeker 2 van de varianties gelijk

#### Voorbeeld

R Square =  $R^2$

Adjusted R Square =  $R^2_a$

**Stap 1:** Globale F-toets

Significantie = 0,000 < 0,05

→  $H_0$  verwerpen dus bruikbaar

**Stap 2:** Hoe goed is het model?

→  $R^2$ : 89,2% wordt verklaard door opgenomen x variabelen

Residual: niet verklaarde deel SSE

Regression: verklaarde deel

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,945 <sup>a</sup>	,892	,885	133,485

a. Predictors: (Constant), Aantal bieders, Ouderdom

**ANOVA<sup>a</sup>**

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4283062.96	2	2141531.48	120.188	.000 <sup>b</sup>
	Residual	516726.540	29	17818.157		
	Total	4799789.50	31			

a. Dependent Variable: Veilingsprijs

b. Predictors: (Constant), Aantal bieders, Ouderdom



#### 4. Schatten en voorspellen

- Schatten: analyse van de coëfficiënten
- Voorspellen: puntschatting (waarden in formule invullen):

Om te schatten moet je realistische gegevens gebruiken → je kan geen waarden invullen die buiten de reikwijdte van de gegeven waarden vallen

Voorspel de veilingprijs voor een:

- klok van 150 jaar oud met 10 bidders →

$$\hat{y} = -1338,95 + 12,74 \cdot 150 + 85,95 \cdot 10 = 1431,55$$

- klok van 50 jaar oud met 2 bidders → geen realistische waarden (enkel gegevens over klokken tussen 100-190j en minstens 5 bidders)

#### 5. Modellen met interactie

Toevoeging: de relatie van Y en x hangt nu wél af van de waarde van de andere x variabele → in SPSS een extra term toevoegen door een vermenigvuldiging te doen van 2 x-waarden en een extra beta.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

$\beta_1 + \beta_3 x_2$ : de verandering in  $E(y)$  voorstelt als  $x_1$  met één eenheid toeneemt ( $x_2$  constant)

$\beta_2 + \beta_3 x_1$ : de verandering in  $E(y)$  voorstelt als  $x_2$  met één eenheid toeneemt ( $x_1$  constant)

Voorbeeld 2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

met:

- $y$  = verkoopprijs van een antieke klok (dollar)
- $x_1$  = ouderdom
- $x_2$  = aantal bidders
- $x_1 x_2$  = interactieterm

1<sup>e</sup> graf: eerste ordemodel: X-variabelen zijn onafhankelijk van elkaar

2<sup>e</sup> graf: effect als  $x_1$  met 1 eenheid

toeneemt hangt af van waarde  $x_2$  → als je meerdere bidders hebt, zal de prijs stijgen bij een klok die 1j ouder is

#### Output

P-waarde  $0,000 < 0,05 \rightarrow H_a \rightarrow$  bruikbaar

$R^2$ : we verklaren 95,4% van  $y$  door  $x$ 'en

Adjusted  $R^2$  is hoger dan vorige model

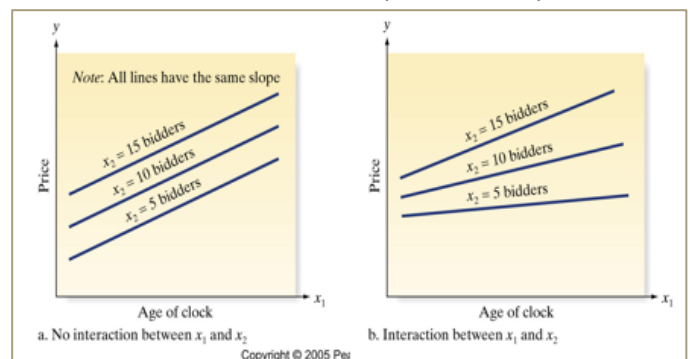
→ grotere verklaringskracht

(Hoge  $R^2$  is niet zo realistisch)

Ouderdom:  $0,669 > 0,05 \rightarrow$  niet significant

Bidders:  $0,004 < 0,05 \rightarrow$  significant

Interactieterm:  $0,000 < 0,05 \rightarrow$  significant



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.977 <sup>a</sup>	.954	.949	88.915

a. Predictors: (Constant), Interactieterm, Ouderdom, Aantal bidders

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4578427.37	3	1526142.46	193.041	.000 <sup>b</sup>
	Residual	221362.133	28	7905.790		
	Total	4799789.50	31			

a. Dependent Variable: Veilingprijs

b. Predictors: (Constant), Interactieterm, Ouderdom, Aantal bidders

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients	Standardized Coefficients	t	Sig.
1	(Constant)	320.458		1.086	.287
	Ouderdom	.878	.061	.432	.669
	Aantal bidders	-93.265	-.673	-3.120	.004
	Interactieterm	1.298	1.369	6.112	.000

a. Dependent Variable: Veilingprijs

significant interactie-effect

Verandering in de verwachte veilingprijs  $E(y)$  van een 150 jaar oude klok voor elke bidder:

$$\frac{\partial E(y)}{\partial x_2} = \hat{\beta}_2 + \hat{\beta}_3 x_1 = -93,265 + 1,298 \cdot 150 = 101,435$$

Ouderdom = niet significant, maar je kan het niet uit model halen door interactieterm

→ stel:  $x_1$  is constant op 150j

Met hoeveel gaat  $y$  toenemen als we een extra bidder hebben?

→ afleiden naar  $x_2$  ( $x_2$  = constant en verdwijnt uit formule)

#### Opletten met interpretatie van Unstandardized Coefficients B!

Invloed van aantal bidders op model zou -93,265 zijn, maar je moet volledig model mét interactieterm bekijken! → veilingprijs zal met 101 euro toenemen

#### Resultaten:

F-toets: het model is bruikbaar

T-toets: interactie-model is significant

$R^2_a$ : hoger dan bij vorig model → verklaringskracht ↑

### 6. Kwadratische en andere hogere-ordemodellen

We zullen krommingen toestaan door met kwadraten te gaan werken (maar blijft lineaire regressie!). We kunnen enkel kwadrateren bij schaalvariabelen.

1. Een kwadratisch (tweede-orde) model met **1 kwantitatieve** verklarende variabele  
→ We gaan een extra term toevoegen waarbij we kwadrateren.

$$E(y) = b_0 + b_1 x + b_2 x^2$$

- $\beta_0$ : snijpunt van kromme met y-as
- $\beta_1$ : verschuivingsparameter
  - $\beta_1 < 0$ : dalende functie
  - $\beta_1 > 0$ : stijgende functie
- $\beta_2$ : mate van de kromming
  - $\beta_2 < 0$ : concave functie
  - $\beta_2 > 0$ : convexe functie

Opnieuw moeten we testen of het model bruikbaar is.

→ zie voorbeeld volgende pagina!

2. Een volledig tweede-ordemodel met **2 kwantitatieve** verklarende variabelen:

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

Volledig model wordt niet aangeraden.

→ Je moet een **keuze** maken: een interactie-effect toevoegen of kwadrateren

→ beste keuze hangt af van de hypothesen die je formuleert

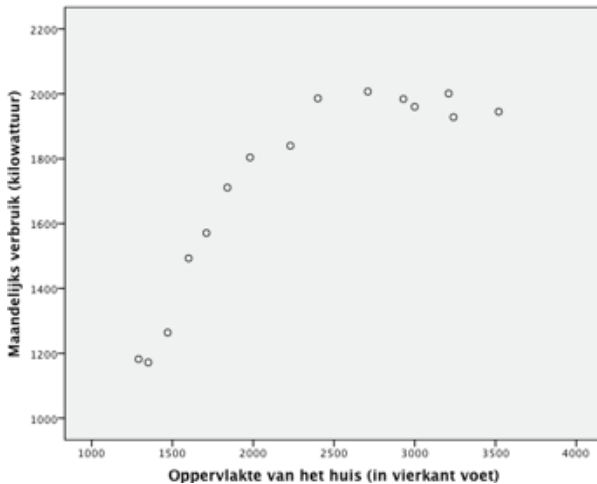
Kwadratisch: als je denkt dat je verband niet rechtlijnig is

Interactie: als je denkt dat het effect van de andere variabelen op  $y$  verschillend is

### Voorbeeld: 1 kwantitatieve variabele

nieuwe variabele:  $x_1^2$  oppervlakte kwadraat

#### Stap 1: spreidingsdiagram



#### Stap 2: bruikbaarheid model

Is het model bruikbaar? ( $\alpha = 0,01$ )

Sig = p-waarde = 0,000 < 0,01 → model bruikbaar

Verklaringskracht: hoog → 97,6%

#### Voorbeeld 4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon$$

met:

- $y$  = maandelijks verbruik (kilowattuur)
- $x_1$  = oppervlakte van het huis (square feet)

Relatie tussen verbruik en oppervlakte  
Verbruik zal toenemen, maar eens je een bepaalde grootte van huis hebt bereikt, zal er niet veel effect meer zijn.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.988 <sup>a</sup>	.976	.972	51.575

a. Predictors: (Constant), Oppervlakte kwadraat, Oppervlakte van het huis (in vierkant voet)

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1314148.55	2	657074.276	247.022	.000 <sup>b</sup>
	Residual	31919.847	12	2659.987		
	Total	1346068.40	14			

a. Dependent Variable: Maandelijks verbruik (kilowattuur)

b. Predictors: (Constant), Oppervlakte kwadraat, Oppervlakte van het huis (in vierkant voet)

#### Stap 3: geschatte coëfficiënten + is er voldoende bewijs voor een concave functie? ( $\alpha = 0,01$ )

Alle coëfficiënten zijn significant: 0,000 < 0,01

$\beta_0 = -838.733$ ;  $\beta_1 = 1.993$ ;  $\beta_2 = -0.000347$

Effect:

- Positieve coefficient bij 'oppervlakte'
- Negatieve coefficient bij 'oppervlakte kwadraat'

→ stijgende concave functie

#### Stap 4: voorspellingsvergelijking + grafisch

$$Y = -838,73 + 1,993 * \text{oppervlakte} - 0,000347 * \text{oppervlakte}^2$$

! geen voorspellingen doen buiten het bereik!

#### Stap 5: Interpreteer de schattingen (zie ook coëfficiënten)

$\beta_1$  is niet meer de helling, met een kwadratische term heeft dit geen zinvolle interpretatie. Voor huizen groter dan 3520m<sup>2</sup> kan men geen schattingen maken omdat deze niet in de steekproef zijn opgenomen.

Coefficients <sup>a</sup>						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-838.733	171.444		-4.892	.000
	Oppervlakte van het huis (in vierkant voet)	1.993	.157	4.869	12.719	.000
	Oppervlakte kwadraat	-0.000347	.000	-4.026	-10.516	.000

a. Dependent Variable: Maandelijks verbruik (kilowattuur)



## 7. Modellen met kwalitatieve variabelen

Kwalitatieve variabelen (nominaal/ordinaal) kan je niet op een schaal meten

→ **Dummyvariabelen** maken: kwalitatieve variabelen coderen als getallen

- Maximaal (m-1) dummyvariabelen voor m categorieën
- '1' = aanwezigheid van die factor
- '0' = afwezigheid van die factor

<b>Voorbeeld:</b>		
Sociale klasse.	x1	x2
Laag.	0	0
Midden	1	0
Hoog	0	1

→ 3 categorieën = 2 dummy's

→ referentiecategorie moet iets zijn waar je mee kan vergelijken en die je kan interpreteren.

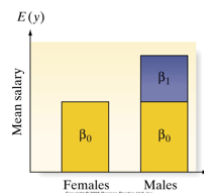
→ Je kan nooit een kwadraat van een dummy nemen, want het heeft geen betekenis. Een interactie kan wel.

Geslacht coderen als dummy-variabele:

$x = 0$  (vrouw)

$x = 1$  (man)

In het model  $E(y) = \beta_0 + \beta_1 x$  geeft  $\beta_1$  het verschil aan tussen het basisniveau en het alternatieve niveau



Verwachte loon vrouw =  $B_0$

Verwachte loon man =  $B_0 + B_1$

Loonverschil is  $B_1$

→ dit toont het effect van waarde 1 (man): wat gebeurt er met je loon als x met 1 toeneemt?

**Voorbeeld 3:**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

met:

- $y$  = schuld in € (bij in gebreken gestelde creditcardgebruikers)
- $x$ : 3 verschillende sociaaleconomische klassen, namelijk lagere klasse, middenklasse, hogere klasse

De verklarende veranderlijke (sociaal-economische klasse) is een kwalitatieve veranderlijke

→ kan 3 waarden aannemen

→ opnemen in model via 2 dummies

$x_1$  = middenklasse (1 voor middenklasse; 0 voor andere waarden)

$x_2$  = hogere klasse (1 voor hogere klasse; 0 voor andere waarden)

→ interpretatie tov referentiecategorie!

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.453 <sup>a</sup>	.205	.146	168.948

a. Predictors: (Constant), Hogere, Midden

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	198772.467	2	99386.233	3.482	.045 <sup>b</sup>
	Residual	770670.900	27	28543.367		
	Total	969443.367	29			

a. Dependent Variable: Schuld in euro

b. Predictors: (Constant), Hogere, Midden

Lagere klasse is hier de referentiecategorie: je gaat veranderingen in het model interpreteren ten opzichte van de lagere klasse.

Bruikbaarheid: f-toets:  $0,045 < 0,05 \rightarrow H_0$  (alle B's = 0) verwerpen dus bruikbaar  
 $R^2$ : 20,5% van de variatie in schuld word verklaard door socio-economische klasse

Coefficients <sup>a</sup>					
Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	t
1	(Constant)	229.600	53.426		4.298
	Midden	80.300	75.556	.211	1.063
	Hogere	198.200	75.556	.520	2.623
a. Dependent Variable: Schuld in euro					

te interpreteren ten opzichte van de referentiecategorie (lagere klasse)

Significatie: midden is niet significant  
 Hogere is wel significant want  $0,014 < 0,05$   
 $B_0 = 229,6$  is schuld lagere klasse  
 Interpretatie: Middenklasse verschilt niet significant met de lagere klasse  
 Hogere klasse heeft een schuld van 198,2 hoger dan die van de lagere klasse

## 8. Modellen met zowel kwantitatieve als kwalitatieve variabelen

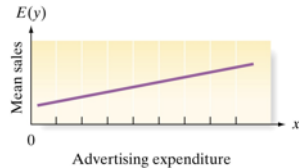
Complexe modellen: je gaat verschillende dingen gaan combineren.

Voorbeelden 5a, 5b en 5c:

- $y$  = maandelijkse omzet
- $x_1$  = maandelijkse reclame-uitgaven
- kwalitatieve veranderlijke: reclamemediã (krant, radio en televisie)
  - $x_2 = 1$  (bij radio), 0 (andere)
  - $x_3 = 1$  (bij televisie), 0 (andere)
  - het medium 'krant' is dus het basisniveau

Eerste-orde model met één kwantitatieve variabele  $x_1$ :

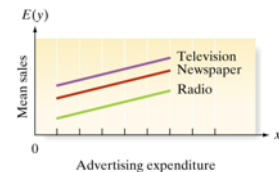
$$E(y) = \beta_0 + \beta_1 x_1$$



Hoe meer uitgaven, hoe hoger de omzet

Toevoegen van de kwalitatieve veranderlijke (zonder interactie):

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

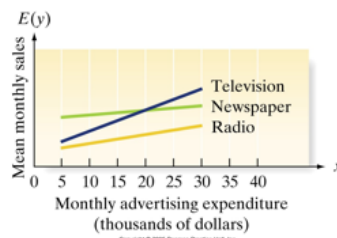


Toevoegen van dummy-variabelen geeft parallelle lijnen  
Snijpunt is verschillend maar helling blijft gelijk.

Bij toevoeging van de interactie-term:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Main effect,  $x_1$       Main effect  $x_2$  and  $x_3$       Interaction



Er zijn verschillen tussen opbrengsten van verschillende mediakanalen

B1 is positief

B2 is negatief (radio slechter dan krant)

B3 is positief (tv beter dan krant)

→ bij een kleine campagne kies je krant, bij een grote campagne kies je beter voor tv

## 9. Residuanalyse: controle van de modelaannames

Residu is verschil tussen wat je waarneemt en wat je voorspelt. Regressie-residu: het verschil tussen een waargenomen  $y$ -waarde en de bijhorende voorspelde waarde.

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Residuanalyse in stappen:

1. Verwachting van  $\varepsilon = 0$  (modelspecificatie)
2. Constante variantie
3. Toevallige afwijkingen normaal verdeeld + uitschieters (outliers) identificeren
4. Toevallige afwijkingen onafhankelijk  
→ soms probleem dat opeenvolgende afwijkingen correleren met elkaar (bij tijdsreeksen)

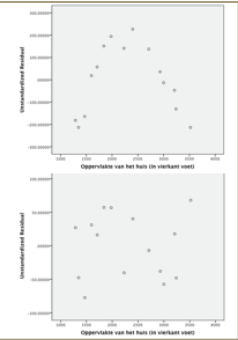
## Modelspecificatie

Je krijgt een plot van uw residu tov een x variabele. Patroon in residuen wijst op ongelijke varianties. Residu moet willekeurig zijn! Als dit niet zo is moet men bvb een kwadratische term toevoegen.

## Modelspecificatie

Bovenste plot toont dat de residuen niet willekeurig verdeeld zijn.

Tweede plot toont residuen die nu wel willekeurig verdeeld zijn (door toevoegen van een 2de-orde term)



## Outliers identificeren

Een plot van de residuen kan aanduiden of er outliers aanwezig zijn in de steekproef.

### Outlier: residu > 3s

De waarneming kan dan eventueel uit de steekproef worden genomen, of er kunnen 2 modellen geschat worden.

### Casewise Diagnostics<sup>a</sup>

Case Number	Std. Residual	Veilingsprijs	Predicted Value	Residual
17	-3,667	1131	1859,60	-728,604

a. Dependent Variable: Veilingsprijs

Waargenomen: 1131

Voorspelde 1859,60

Residu van 728,604

Gestandaardiseerd is dit 3,667

→ uitschieter

## Een normale verdeling ε nagaan via:

- Histogram
- normal probability plot
- in SPSS volgens de **Kolmogorov-Smirnov toets**

H<sub>0</sub>: de variabele is normaal verdeeld

H<sub>a</sub>: de variabele is niet normaal verdeeld

→ kies de waargenomen residuen als bestudeerde variabele

→ p-waarde vergelijken met  $\alpha$  om tot besluit te komen



## Output

### One-Sample Kolmogorov-Smirnov Test

	Standardized Residual
N	15
Normal Parameters <sup>a,b</sup>	
Mean	,0000000
Std. Deviation	,92582010
Most Extreme Differences	
Absolute	,186
Positive	,186
Negative	-,167
Test Statistic	,186
Asymp. Sig. (2-tailed)	,174 <sup>c</sup>

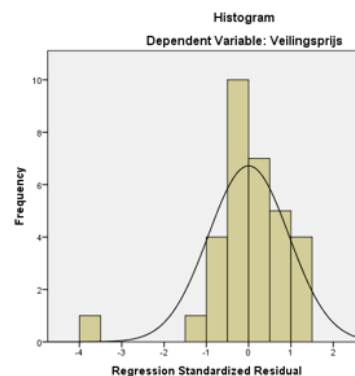
a. Test distribution is Normal.

**P = 0,174 >  $\alpha$  = 0,05**

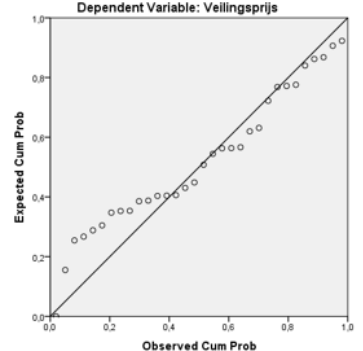
Ho aanvaarden

De verdeling van de waargenomen residuen is normaal verdeeld

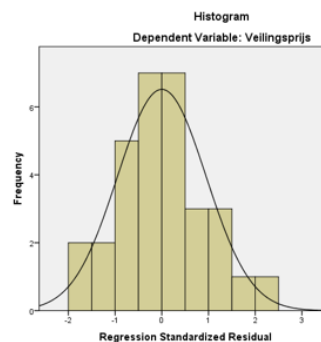
### met outlier



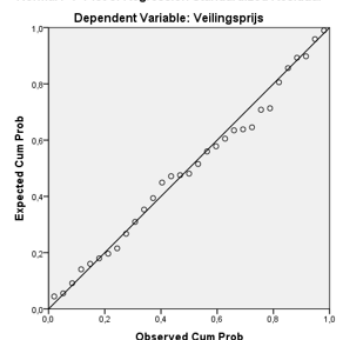
### Normal P-P Plot of Regression Standardized Residual



### zonder outlier



### Normal P-P Plot of Regression Standardized Residual



## 10. Enkele valkuilen

**Schatbaarheid:** het aantal niveaus van waargenomen x-waarden mag niet minder zijn dan het aantal te schatten  $\beta_i$ 's

Je moet zorgen dat je genoeg cases hebt. Je hebt voldoende info nodig vooraleer je een nuttige regressie kan doen.

### **Multicollineariteit:**

wanneer 2 of meerdere onafhankelijke variabelen sterk gecorreleerd zijn, leidt tot verwarrende en misleidende resultaten (bv. tekens van de geschatte  $\beta$ -waarden)

- Identificatie: je moet deze identificeren door:
  - Correlaties bepalen tussen alle x'en (correlatiematrix)
  - Statistische toetsen (VIF, Tolerance): bekijken in basismodel (dwz model zonder interactie of kwadratische term)!
- Oplossingen:
  - Een van de correlerende variabelen weglaten uit het model. Je gebruikt dit om te kijken of de keuze van de controlevariabele wel goed is.
  - Enkel naar de globale F-toets kijken en niet naar t-toetsen

→ **Tolerantie** voor de veranderlijke  $x_i$ :

$$\text{Tol}_i = 1 - R_i^{*2}$$

- Waarbij  $R_i^{*2}$  de determinatiecoëfficiënt is voor de voorspelling van de (verklarende) veranderlijke  $x_i$  door de andere verklarende veranderlijken

**Lage tolerantie-waarden wijzen dus op een variabele die weinig toevoegt aan het model. Drempelwaarde:  $\text{Tol}_i < 0,1$ .**

Als dus 90% van de x verklaard is door de andere X'en, kan je die er beter uitlaten

$R^2$  is hoeveel % van u variatie in y wordt verklaard door opgenomen door u x variabele  
 $R^{*2}$  hoeveel % van de variatie in de ene x wordt verklaard door u andere X'en

→ **Variance Inflation Factor (VIF)** voor de veranderlijke  $x_i$ :

$$\text{VIF}_i = 1 / \text{Tol}_i$$

**Een hoge VIF-waarde wijst op een variabele die weinig toevoegt aan het model. Drempelwaarde:  $\text{VIF} > 10$ .**

Voorspellingen buiten het experimentele gebied: het is gevaarlijk om een regressie-model te gebruiken voor schattingen buiten het bereik van het model.



# PPT: Meervoudige regressie analyse (SPSS)

(enkel aanvullende info bij theorie of specifieke toepassingen, geen herhaling)

## 1. Constructie model

Hoe neem ik verschillende types variabelen op in een regressiemodel?

- Kwantitatieve variabelen ("scale"): gewoon opnemen
- Ordinale variabelen ("ordinal"): 1 of meerdere dummy's maken
- Nominale variabelen ("nominal"): 1 of meerdere dummy's maken

### Voorbeeld: meerdere dummies

Beschouw variabele *regionbe* (155) -> deze variabele kan **3 waarden aannemen**:  
1: Flemish; 2: Walloon; 3: Brussels

Hoe neem je deze variabele op in een regressiemodel? Antwoord: **2 dummies**

	Dummy Vlaanderen ( $x_1$ )	Dummy Wallonië ( $x_2$ )
Vlaanderen	1	0
Wallonië	0	1
Brussel	0	0

Referentie: Brussel -> in alle dummies 0

$\beta_1 x_1 + \beta_2 x_2$  -> alle mogelijke waarden van de oorspronkelijke variabele *regionbe* worden meegenomen in het model.

Interpretatie:

$\beta_1$ : effect van Vlaanderen tov referentie Brussel

$\beta_2$ : effect van Wallonië tov referentie Brussel

## Uitbreidingen lineaire regressie

	Afhankelijke var (y)	Onafhankelijke var (x'en)
Logistische regressie	dichotome (nom/ord) variabele	scale var OF dummy var
Ordinal Logit model	ordinale variabele	
Multinomial logit model	nominale variabele	

## 2. - 3. - 4. Eenvoudig voorbeeld – interpretaties

### Lineaire regressie uitvoeren in SPSS -> output

Wat willen we onderzoeken?

- Beschrijvende statistieken: eventuele fouten in hercoderingen opsporen
- Bruikbaarheid van het model toetsen via globale F-toets + verklaringskracht onderzoeken via  $R^2$
- Analyse van de coëfficiënten
  - Significantie van coëfficiënten: individuele t-toetsen
  - Effect per variabele onderzoeken: ongestandaardiseerde beta's
  - Vergelijken over variabelen: gestandaardiseerde beta's
- Analyse van de werkhypothesen
  - Multicollineariteit controleren: VIF en Tolerance
  - Outliers detecteren
  - Normaliteit van de afwijkingen nagaan (residuen) via histogram en normal probability plot



### Voorbeeld 1:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

met:

- $y$  = totaal aantal werkuren per week (inclusief overuren) → schaalvariabele
- $x_1$  = leeftijd → schaalvariabele
- $x_2$  = geslacht  
→ hercoderen naar 0/1 dummy variabele gender  
(0 'female' = referentiecategorie en 1 'male')
- $x_3$  = aantal jaren opleiding → schaalvariabele

## 1) BESCHRIJVENDE STATISTIEKEN

Dummy gemaakt voor geslacht:

0 = vrouw (referentie)

1 = man

(Standaardafwijkingen uit deze tabel hebben we later nodig bij gestandaardiseerde beta's)

Descriptive Statistics

	Mean	Std. Deviation	N
Total hours normally worked per week in main job overtime included	38,50	16,476	1679
Age of respondent, calculated	48,43	17,885	1679
dummy gender	,5045	,50013	1679
Years of full-time education completed	13,19	3,817	1679

## 2) BRUIKBAARHEID VAN HET MODEL

- Globale F-toets
- $R^2$  en  $R^2_a$
- Standaardafwijking van het model

### Globale F-toets

Minimale voorwaarde  
bruikbaarheid

$H_0: \beta_1 = \beta_2 = \beta_3 = 0$

$H_a$ : ten minste 1  $\beta$  is verschillend van 0

$P(2\text{-zijdig}) = 0,000 < 0,05$

→  $H_0$  verwerpen

→ Het model is bruikbaar: ten minste 1 x veranderlijke heeft een significant verband met  $y$

Model Summary<sup>b</sup>

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,303 <sup>a</sup>	,092	,090	15,717

a. Predictors: (Constant), Years of full-time education completed, dummy gender, Age of respondent, calculated

b. Dependent Variable: Total hours normally worked per week in main job overtime included

ANOVA<sup>a</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	41738,275	3	13912,758	56,318	,000 <sup>b</sup>
	Residual	413789,457	1675	247,038		
	Total	455527,732	1678			

a. Dependent Variable: Total hours normally worked per week in main job overtime included

b. Predictors: (Constant), Years of full-time education completed, dummy gender, Age of respondent, calculated

$F = \text{Mean Square Regression} /$

$\text{Mean Square Residual} = 13912,758 / 247,038 = 56,318$

→ Als  $F=1$  dan zijn de varianties perfect gelijk. Hoe verder  $F$  is van 1, hoe meer waarschijnlijk dat de varianties verschillen.

**Verklaringskracht:  $R^2$  en  $R^2_a$**  (zie formules in theorie gedeelte)

*Hoe goed past het model bij de gegevens?*

$R^2 = R \text{ Square} = \text{verklaarde variantie} / \text{totale variantie}$

$$= 41738,275 / 455527,732 = 0,0916$$

→ 9,2% van de variatie in y (het aantal werkuren per week) wordt verklaard door de opgenomen x-variabelen (leeftijd, geslacht, opleiding)

### **Standaardafwijking van het model**

*Hoe nauwkeurig is het model om voorspellingen te doen?*

$$s^2 = 413789,457 / 1675 = 247,038$$

$$s = \text{Std. Error of the Estimate} = \sqrt{247,038} = 15,717$$

→ hoe kleiner s, hoe accurater de voorspellingen van het model

Regression: de variantie die verklaard wordt door de x-variabelen

→ df = aantal coëfficiënten - 1 (hier 4 coëfficiënten  $\beta_0, \beta_1, \beta_2, \beta_3$ )

Residual: de variantie die verklaard wordt door de y variabele

Total: regression + residual

→ df = aantal waarnemingen - 1 (hier N = 1679)

### **3) ANALYSE VAN DE COËFFICIËNTEN**

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 $\beta_0$ (Constant)	21,845	2,026		10,781	,000		
$\beta_1$ Age of respondent, calculated	,168	,022	,182	7,505	,000	,922	1,084
$\beta_2$ dummy gender	7,779	,769	,236	10,119	,000	,996	1,004
$\beta_3$ Years of full-time education completed	,349	,105	,081	3,337	,001	,921	1,086

a. Dependent Variable: Total hours normally worked per week in main job overtime included

#### **- Individuele t-toetsen: zijn de coëfficiënten significant?**

Bv voor leeftijd x1:

Ho:  $\beta_1 = 0$  (de coëfficiënt voor leeftijd is gelijk aan 0)

Ha:  $\beta_1 \neq 0$  (de coëfficiënt voor leeftijd verschilt significant van 0)

P = 0,000 < 0,05 → Ho verwerpen → Leeftijd heeft een significante invloed op y

Bij elke x-variabele geldt P-waarde < 0,05 → elke coëfficiënt heeft een significante invloed op y.

- **Ongestandaardiseerde coëfficiënt: effect per variabele**

Via unstandardized coefficients B:

$$y_e = 21,845 + 0,168 * \text{leeftijd} + 7,779 * \text{gender} + 0,349 * \text{aantal jaren opleiding}$$

➤ Effect van ongestandaardiseerde coëfficiënt:

Wat is het effect op y van een toename van x met 1 eenheid (als alle andere x'en constant blijven)?

→ effecten zijn onafhankelijk van de waarde van de andere variabelen

→ opletten met interpretatie bij dummy variabelen! (\*)

→ Let op: bij een variabele die niet significant is kan je geen effect bespreken (gezien de impact niet significant verschillend is van 0)

➤ Bv voor leeftijd: 1 jaar ouder is gelijk aan 0,168u meer werken

➤ Bv voor geslacht(\*): Een man zal 7,779u meer werken dan een vrouw

→ spreek niet over een toename van geslacht met één eenheid!

- **Gestandaardiseerde beta's: vergelijken over variabelen**

Geven aan welke x-variabele de grootste impact heeft op y

→ omdat beta's rekening houden met de verschillen in eenheden tussen de onafhankelijke veranderlijken.

De gestandaardiseerde  $\beta$ 's worden als volgt berekend:

$$\text{beta}_i = \beta_i * (s_{xi} / s_y)$$

→  $\beta_i$  = ongestandaardiseerde  $\beta$  van de overeenkomende x

→ voor standaardafwijkingen zie 'descriptive statistics' (1<sup>e</sup> tabel)

Bv voor leeftijd x1:

$$\beta_1 = 0,168 * (17,885 / 16,476) = 0,182$$

Voorspel het aantal uren werken voor een: vrouw van 40 jaar met 15 jaar opleiding

$$y_e = 21,845 + 0,168 * 40 + 7,779 * 0 + 0,349 * 15 = 33,8u$$

Let op: de ingevulde waarden moeten tot de reikwijdte van het model behoren!

→ Man van 200 jaar: geen realistische waarden

#### 4) ANALYSE VAN DE WERKHYPOTHESEN

##### – Multicollineariteit:

wanneer 2 of meerdere onafhankelijke variabelen sterk gecorreleerd zijn

→ leidt tot verwarrende en misleidende resultaten (bv. tekens geschatte  $\beta$ -waarden)

- Identificatie
  - correlaties bepalen tussen alle x'en (correlatiematrix)
  - statistische toetsen (VIF, Tolerance)
- Oplossingen
  - één van de correlerende variabelen weglaten uit het model

#### Correlatiematrix

		Correlations			
		Total hours normally worked per week in main job overtime included	Age of respondent, calculated	dummy gender	Years of full-time education completed
Pearson Correlation	Total hours normally worked per week in main job overtime included	1,000	,167	,245	,042
	Age of respondent, calculated	,167	1,000	,029	-,275
	dummy gender	,245	,029	1,000	,047
	Years of full-time education completed	,042	-,275	,047	1,000
Sig. (1-tailed)	Total hours normally worked per week in main job overtime included		,000	,000	,043
	Age of respondent, calculated	,000		,117	,000
	dummy gender	,000	,117		,026
	Years of full-time education completed	,043	,000	,026	

#### Statistische toetsen

##### 1) Tolerantie voor de veranderlijke $x_i$ :

$$\text{Tol}_i = 1 - R_i^{*2}$$

waarbij  $R_i^{*2}$  de determinatiecoëfficiënt is voor de voorspelling van de (verklarende) veranderlijke  $x_i$  door de andere verklarende veranderlijken

→ lage tolerantie-waarden wijzen dus op een variabele die weinig toevoegt aan het model

→ drempelwaarde: **Tol  $< 0,1$**

##### 2) Variance Inflation Factor (VIF) voor de veranderlijke $x_i$ :

$$\text{VIF}_i = 1 / \text{Tol}_i$$

→ hoge VIF-waarden wijzen dus op een variabele die weinig toevoegt aan het model

→ drempelwaarde: **VIF  $> 10$**

#### Probleem als Tol $< 0,1$ of VIF $> 10$

- oplossing 1: variabele weglaten uit het model
  - Variabele met laagste gestandaardiseerde beta
- oplossing 2: transformatie van de variabelen (komt niet aan bod)

!Opgelet: altijd nagaan in basismodel (dwz model zonder interactie of kwadratische termen)!

Hier geldt voor alle x-variabelen: Tol  $> 0,1$  en VIF  $< 10$  → onafhankelijke variabelen zijn niet onderling gecorreleerd & elke variabele voegt iets toe aan het model. We hoeven geen variabele uit het model te halen.

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
1 (Constant)	21,845	2,026		10,781	,000		
Age of respondent, calculated	,168	,022	,182	7,505	,000	,922	1,084
dummy gender	7,779	,769	,236	10,119	,000	,996	1,004
Years of full-time education completed	,349	,105	,081	3,337	,001	,921	1,086

a. Dependent Variable: Total hours normally worked per week in main job overtime included

– **Outliers identificeren:** **Outlier: residu > 3s**

→ Nagaan of het hier eventueel om een fout gaat.

→ De waarneming kan dan eventueel uit de steekproef worden genomen, of er kunnen 2 modellen geschat worden.

Bij dit model zijn er 13 outliers – voornamelijk bij mensen die of niet werken of net heel veel...

→ Nagaan of er geen fouten gemaakt zijn bij de codering!

→ dit zeker vermelden

**Casewise Diagnostics<sup>a</sup>**

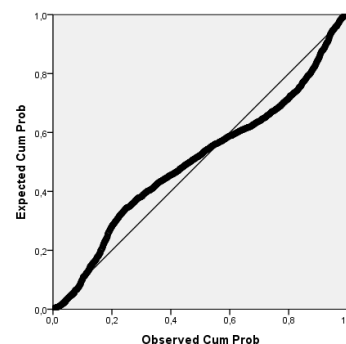
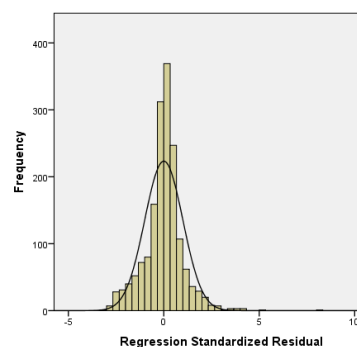
Case Number	Std. Residual	Total hours normally worked per week in main job overtime included	Predicted Value	Residual
690	4,104	105	40,50	64,505
742	3,686	100	42,06	57,939
817	5,229	126	43,82	82,181
824	4,083	105	40,83	64,170
953	4,115	100	35,33	64,670
992	3,443	100	45,89	54,112
1264	3,224	96	45,33	50,669
1427	3,488	90	35,17	54,826
1478	-3,059	0	48,08	-48,082
1636	3,696	100	41,91	58,093
1802	8,143	168	40,02	127,982
1856	3,499	90	35,01	54,994
1866	3,983	100	37,40	62,604

a. Dependent Variable: Total hours normally worked per week in main job overtime included

– **Normale verdeling van de afwijkingen**

➤ Grafisch:

- Histogram
- Normal P-P Plot



➤ Kolmogorov-Smirnov toets (SPSS):

De meervoudige regressie-analyse is robuust tegen afwijkingen van de normaliteit: deze afwijkingen hebben weinig effect op de analyses (centrale limietstelling).

$H_0$ : de residuen zijn normaal verdeeld

$H_a$ : de residuen zijn niet normaal verdeeld

→ p-waarde >  $\alpha$  : normale verdeling van de residuen

Hier:  $p = 0,000 < 0,05$

→  $H_0$  verwerpen

→ residuen zijn niet normaal verdeeld

**One-Sample Kolmogorov-Smirnov Test**

		Standardized Residual
N		1679
Normal Parameters <sup>a,b</sup>	Mean	,0000000
	Std. Deviation	,99910568
Most Extreme Differences	Absolute	,095
	Positive	,086
	Negative	-,095
Test Statistic		,095
Asymp. Sig. (2-tailed)		,000 <sup>c</sup>

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

(Voor resterende delen lees je gewoon eens de slides, er wordt niks nieuws gezegd dat hier niet vermeld is maar dit zijn goede oefeningen.)

### 3. Interactie-model

### 4. Kwadratisch model

### 5. Uitgebreid voorbeeld – rapportering

### 6. Alternatieve settings

We nemen het product tussen twee verschillende onafhankelijke variabelen op indien we rekening willen houden met de afhankelijkheid tussen deze twee variabelen

## PPT: Wetenschappelijk rapporteren

Focus is je hypothese onderzoeken

- Statistiek is een hulpmiddel
- Geen handleiding schrijven (bvb redenering nulhypothese verwerpen niet uitschrijven )
- Kom tot een inhoudelijke conclusie (is je resultaat in overeenstemming met je vooropgestelde hypothese)
  
- Geef informatie over de variabele die je gaat gebruiken voor je analyse (zowel de vraag als de antwoorden).
- Geef geen SPSS-output. Maak zelf een tabel en geef een nummer en titel aan tabellen en figuren EN verwijs ernaar in je tekst.
- Hercodeer variabelen enkel indien dit noodzakelijk is om de specifieke analyse uit te voeren. Het is altijd beter om te werken met meer variatie.
- Bij t-toets: tabel met gemiddelden weergeven en deze bespreken. Geen voorbeeld van t-toets uitwerken maar inhoudelijk bespreken welke variabelen wel en niet significant zijn.
- Bij Chi-kwadraat toets: kruistabel opnemen en bespreken.
- Doel: inhoudelijk rapporteren, minder statistisch. Aangeven of er significant verband is en wat je kan afleiden uit de tabel over dit verband (niet achterliggende nul- en alternatieve hypothese geven).
- In bachelor- en masterproef worden hypotheses gemotiveerd vanuit de literatuur en koppel je ook terug naar de literatuur (overeenstemming of niet, zo niet wat zou de oorzaak kunnen zijn: ander land/periode/operationalisering).
- Controlevariabelen kies je ifv van literatuur en beschikbaarheid data (denk logisch na ipv het zo gemakkelijk mogelijk te houden). Durf ook een kwadratisch of interactiemodel aan.
- Zeker ingaan op de bespreking van het effect voor de variabele uit je hypothese.
- Koppel terug naar je bivariate analyses.
- Ga ook kort in op outliers, multi-collineariteit en normaliteit van de residuen.  
Aandachtspunt: Test werkhypothese adhv Kolmogorov-Smirnov toets is om na te gaan of residuen normale verdeling hebben niet voor andere variabelen.