Anvendt Statistik
Lektion 9

Variansanalyse (ANOVA)

Undersøge sammenhæng

- Undersøge sammenhænge mellem **kategoriske variable**:
  - χ^2-test i kontingenstabeller
- Undersøge sammenhæng mellem **kontinuerte variable**:
  - Multipel eller simpel lineær regression.
- Undersøge forskellen i **middelværdi for to grupper**

- Denne gang:
- Sammenligne middelværdier i mere end to grupper
  - Metode: Variansanalyse (**AN**alysis **O**f **VA**riance)
  - Eksempel: Er der forskel i middelløn for tre grupper

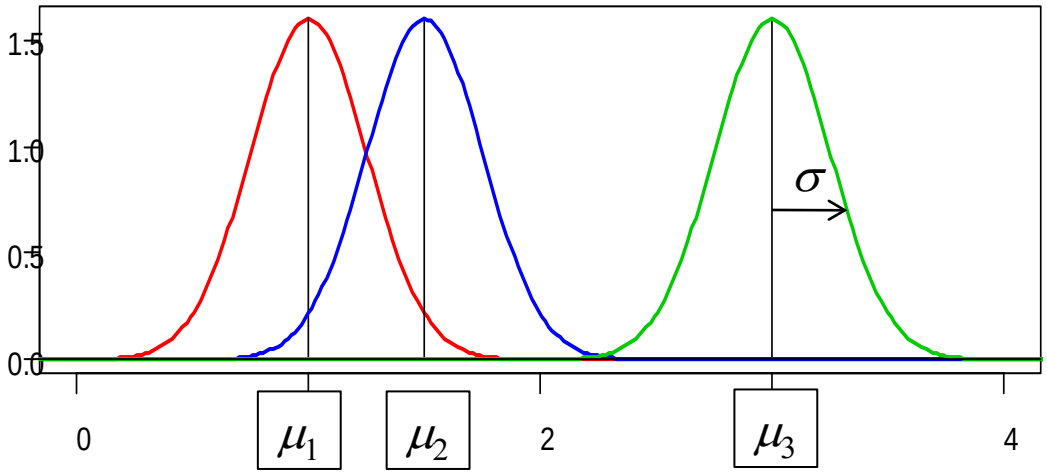
ANOVA: Setup

- Vi har
  - g **grupper**
  - Dvs. hvis vi vil sammenligne tre grupper, så er $g = 3$
- De g grupper har **middelværdierne**
  - $\mu_1, \mu_2, \ldots, \mu_g$
  - Dvs. μ_1 er middelværdi for gruppe 1, osv.

- **Variansanalyse** er et ***F*-test** af
  - $\mathbf{H_0}$: $\mu_1 = \mu_2 = \mu_g$ (ens middelværdier)
  - $\mathbf{H_a}$: Mindst en middelværdi skiller sig ud

Antagelser

- **Antagelser** for at *F*-testet i ANOVA er gyldigt:
  - Hver af de *g* grupper er normalfordelte
  - Standardafvigelsen, σ, er den samme for alle grupper
  - De *g* stikprøver er uafhængige



Hypotese og Fortolkning

- **Variansanalyse** er et ***F*-test** af
  - $\mathbf{H_0}$: $\mu_1 = \mu_2 = \mu_g$ (ens middelværdier)
  - $\mathbf{H_a}$: Mindst en middelværdi skiller sig ud

- **Fortolkning**: Hypoteserne har følgende fortolkning
  - $\mathbf{H_0}$: Ingen effekt af den forklarende variabel
  - $\mathbf{H_a}$: Den forklarende variabel *har* en effekt

- Hvis vi afviser H_0, så kan årsagen fx være at
  - Én gruppe skiller sig ud
  - Alle grupper har forskellige middelværdier

Eksempel: Politisk Ideologi

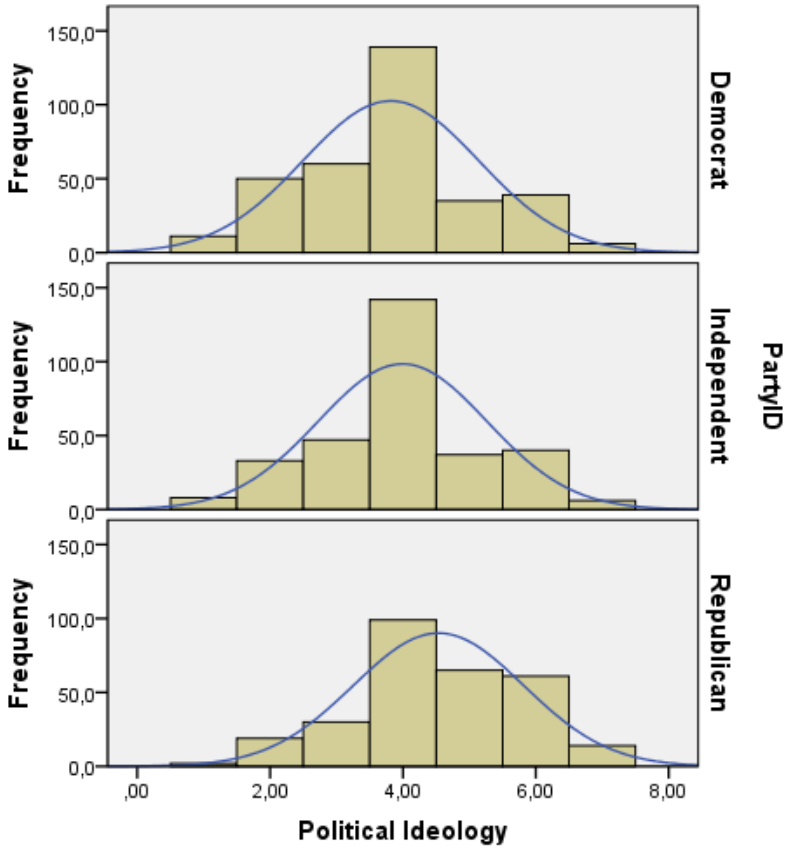
- Hver af 943 personer har angivet:
- **Parti**
  - Demokrat, Uafh., Republikaner
- **Politisk ideologi**
  - Heltal fra 1 til 7

- Opsummering af data:

Political Ideology

PartyID	N	Mean	Std. Deviation
Democrat	340	3,8176	1,32226
Independent	313	3,9936	1,26844
Republican	290	4,5345	1,28359
Total	943	4,0965	1,32597

SPSS: Analyze → Compare Means → Means



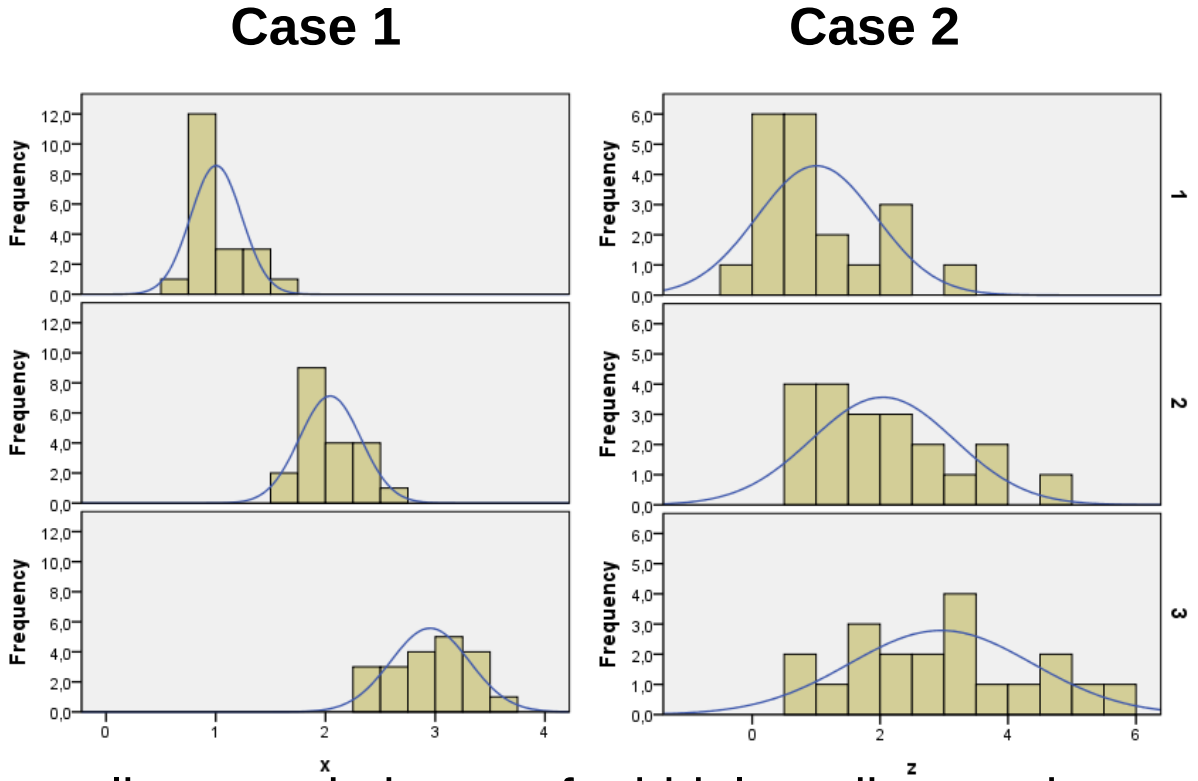
SPSS: Chart builder: Histogram + Groups/Point ID → Rows panel variable

Mærkeligt navn…

- Hvorfor hedder det variansanalyse, når det handler om at sammenligne middelværdier???

- **Case 1**:
- Tydelig forskel i middelværdi!
- **Case 2**:
- Ikke så tydeligt…

- De tre middel-værdier er de samme i begge cases!!



- **Forskellen**: Vi sammenligner variationen af middelværdien med variationen i hver af de tre grupper. Derfor hedder det variansanalyse

F-testet: Forhold af variansestimater

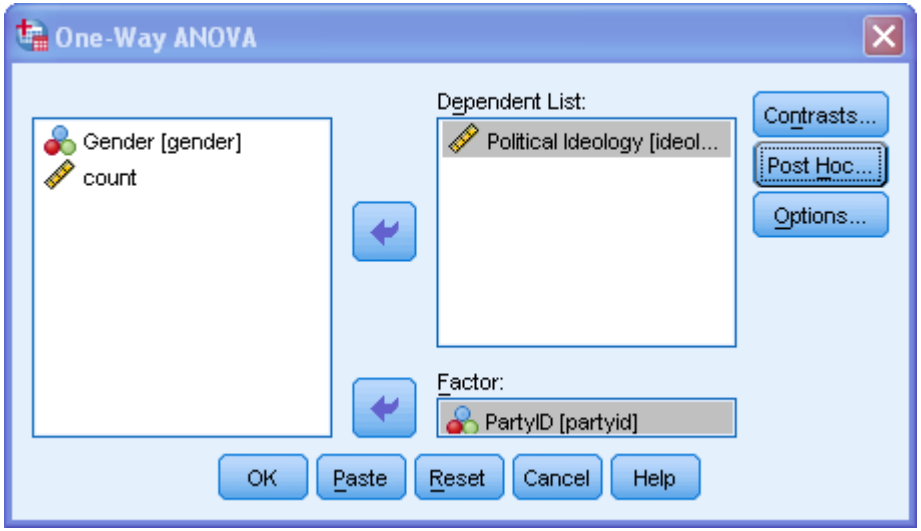
- Notation:
  - $\bar{y}_i$ gennemsnittet i *i*'te gruppe
  - $\bar{\bar{y}}$ gennemsnittet af *alle* data
- *F*-teststørrelsen er

$$F = \frac{\text{Between-groups variansestimat}}{\text{Within-groups variansestimat}}$$

- Variansestimater:
- **Between-groups**: Baseret på variationen i $\bar{y}_i$ 'erne (omkr. $\bar{\bar{y}}$).
  - Er et unbiased estimat af σ^2, *hvis* H_0 er sand.
- **Within-groups**: Baseret på variationen i grupperne.
  - Er *altid* et unbiased estimat af σ^2!
- Hvis H_0 er falsk, har *F* tendens til at være stor.

Eksempel

- **SPSS**: Analyze → Comapre Means → One-Way ANOVA
- H_0 afvises – der *er* en forskel i middelværdierne.



Between-Groups variansestimat

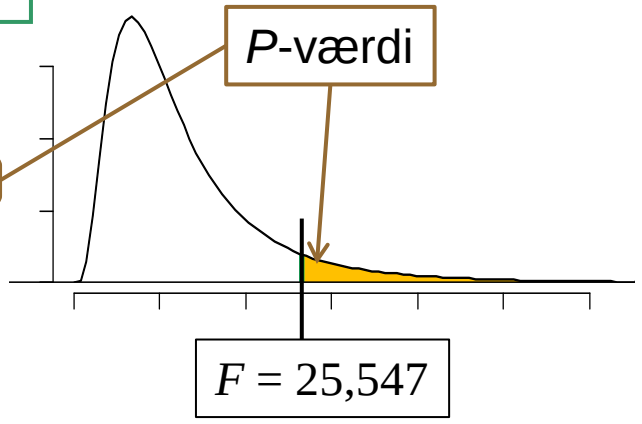
$$F = \frac{42{,}691}{1{,}671} = 25{,}547$$

ANOVA

Political Ideology

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	85,382	2	42,691	25,547	,000
Within Groups	1570,837	940	1,671		
Total	1656,218	942			

Within-Groups variansestimat



Sammenligninger af mange middelværdier

- Antag **vi har afvist H_0**, dvs. middelværdierne er forskellige.
- **Spørgsmål**: Hvilken middelværdi skiller sig ud?
- **Ide**: Udregn konfidensintervaller for forskellen i middelværdi for alle par af middelværdier:

- Et konfidensinterval for $\mu_i - \mu_j$ er

$$\bar{y}_i - \bar{y}_j \pm t \cdot s \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

- t har $df = N - g$ frihedsgrader.

Eksempel

- Find et 95% konfidensinterval for forskellen i middel ideologi for demokrater og republikanere:
- Demokrater: $\bar{y}_1 = 3.82$, $n_1 = 340$.
- Republikanere: $\bar{y}_3 = 4.53$, $n_3 = 290$.

Political Ideology

PartyID	N	Mean	Std. Deviation
Democrat	340	3,8176	1,32226
Independent	313	3,9936	1,26844
Republican	290	4,5345	1,28359
Total	943	4,0965	1,32597

- 95% konfidensinterval for $\mu_3 - \mu_1$:

$$\bar{y}_i - \bar{y}_j \pm t_{0.025} \cdot s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \qquad (0.51; 0.92)$$

- Dvs. vi er 95% sikre på at forskellen er mellem 0.51 og 1.12.

Mange sammenligninger

- Har vi $g = 10$ grupper laver vi $g(1\text{-}g)/2 = 45$ parvise sammenligninger fx vha. 95% konfidensintervaller.
- Hvert konfidensinterval vil *isoleret* set indeholde den sande forskel med 95% sikkerhed.
- Derimod vil de 45 intervaller typisk *ikke* alle **samtidigt** indeholde den sande værdi med 95% sikkerhed!

- **Løsning**: **Bonferroni** sammenligning
  - Antag vi har $g = 4$ grupper, dvs. 6 sammenligninger.
  - I stedet for $(1-\alpha)100\% = 95\%$ konfidensintervaller ($\alpha = 5\%$), så bruger vi $(1-\alpha/6)100\% = 99.2\%$ konfidensintervaller.
  - Dette sikre at konfidensniveauet er *mindst* 95%.

Eksempel: Bonferroni

- Forskellen mellem demokrater og republikanere:
- $g = 3$, dvs. $3(3\text{-}1)/2 = 3$ sammenligninger.
- Så vi skal bruge $\alpha = 0{,}05/3 = 0{,}017$.

$$\bar{y}_i - \bar{y}_j \pm t_{0.0083} \cdot s\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = \qquad (0.47\,;\,0.96)$$

- I SPSS vælger man Bonferroni under 'Post-hoc'

Variansanalyse og Regression

- Vi kan formulere en variansanalyse som en multipel lineær regression!
- Det kræver vi indfører såkaldte **dummy-variable**.
- **Eksempel**: Vi har $g = 3$ grupper
  - Vi indfører to **dummy variable** z_1 og z_2, der indikerer om en observation tilhører hhv. gruppe 1 eller 2.

Obs. grp.	z_1=	z_2=
1	1	0
2	0	1
3	0	0

  - Dvs. for en observation fra gruppe 2 har vi $z_1 = 0$ og $z_2 = 1$.

Regressionsmodel

- Vi kan nu formulere en multipel lineær regressionsmodel:

$$E[y] = \alpha + \beta_1 z_1 + \beta_2 z_2$$

- For **gruppe 1** har vi $z_1 = 1$ og $z_2 = 0$ dvs.

$$E[y] = \alpha + \beta_1 \cdot 1 + \beta_2 \cdot 0 = \alpha + \beta_1 = \mu_1$$

- For **gruppe 2** har vi $z_1 = 0$ og $z_2 = 1$ dvs.

$$E[y] = \alpha + \beta_1 \cdot 0 + \beta_2 \cdot 1 = \alpha + \beta_2 = \mu_2$$

- For **gruppe 3** har vi $z_1 = 0$ og $z_2 = 0$ dvs.

$$E[y] = \alpha + \beta_1 \cdot 0 + \beta_2 \cdot 0 = \alpha = \mu_3$$

Fortolkning

- Vi kan nu formulere en multipel lineær regressionsmodel:

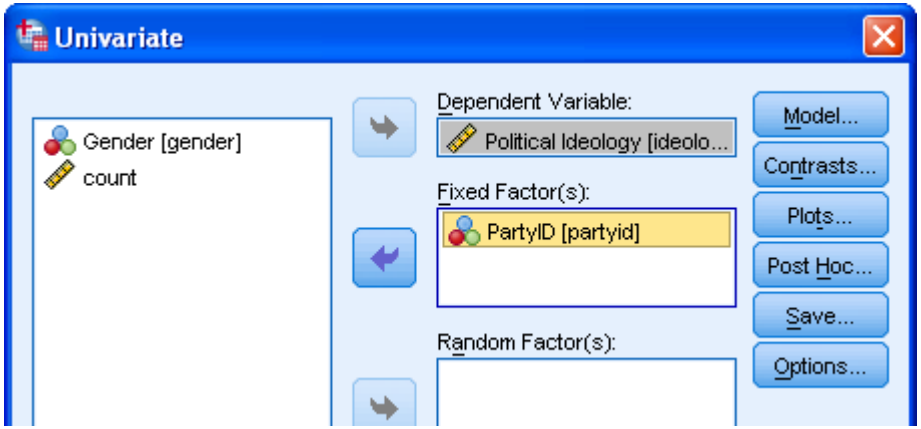
$$E[y] = \alpha + \beta_1 z_1 + \beta_2 z_2$$

Gruppe	z_1=	z_2=	Middelv. for *y*	Fortolkning af β
1	1	0	$\mu_1 = \alpha + \beta_1$	$\beta_1 = \mu_1 - \mu_3$
2	0	1	$\mu_2 = \alpha + \beta_2$	$\beta_2 = \mu_2 - \mu_3$
3	0	0	$\mu_3 = \alpha$	

- α kan fortolkes som middelværdien for gruppe 3 (referencegruppen)
- β_1 og β_2 kan fortolkes som forskelle i middelværdien for hhv. gruppe 1 og 2 i forhold til referencegruppen (gruppe 3)

Estimation

- **SPSS**: Analyze → General Linear Model → Univariate
- Under options vælg 'Parameter estimates'
- Output:



Parameter Estimates

Dependent Variable:Political Ideology

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4,534	,076	59,734	,000	4,386	4,683
[partyid=1,00]	-,717	,103	-6,937	,000	-,920	-,514
[partyid=2,00]	-,541	,105	-5,133	,000	-,748	-,334
[partyid=3,00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

a

b_1

b_2

- **Estimerede model**:

$$\hat{y} = 4.535 - 0.717 \cdot z_1 - 0.541 \cdot z_2$$

- Dvs. den estimerede middelværdi for gruppe 1 er:

$$4.535 - 0.717 \cdot 1 - 0.541 \cdot 0 = 4.535 - 0.717 = 3.818$$

Hypotesetest i Regressionsmodel

- I **multipel lineær regression** udførte vi et *F*-test af hypotesen:
  - H_0: $\beta_1 = \beta_2 = 0$
  - H_a: mindst et $\beta_j \neq 0$
- **Fortolkningen** af H_0: Alle grupper har samme middelværdi.

- Det svarer præcist til *F*-testet i ANOVA
  - H_0: $\mu_1 = \mu_2 = \mu_3$
  - H_a: Mindst et μ_j skiller sig ud.

- Dvs. der er intet tabt ved at bruge regressionsformuleringen.

Hypotesetest i SPSS

- **SPSS**: Analyze → General Linear Model → Univariate

Tests of Between-Subjects Effects

Dependent Variable:Political Ideology

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	85,382[a]	2	42,691	25,547	,000
Intercept	15902,742	1	15902,742	9516,317	,000
partyid	85,382	2	42,691	25,547	,000
Error	1570,837	940	1,671		
Total	17481,000	943			
Corrected Total	1656,218	942			

a. R Squared = ,052 (Adjusted R Squared = ,050)

- **Bemærk**: Resultat er præcist som når vi bruger One-Way ANOVA funktionen i SPSS.

ANOVA

Political Ideology

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	85,382	2	42,691	25,547	,000
Within Groups	1570,837	940	1,671		
Total	1656,218	942			

To-sidet Variansanalyse (Two-Way ANOVA)

- **Indtil nu**: Hvordan middelværdien for én kontinuert variabel (Ideologi) afhænger af én kategorisk variabel (Parti ID): **En-sidet** variansanalyse.

- Vi vil nu se på, hvordan én kontinuert variabel afhænger af to kategorisk variabel
- **Eksempel**:
- **Ideologi** forklaret ved

  Parti ID *og* **køn**
- **SPSS**: Compare Means → Means…
- Tilføj PartyID og Gender i hvert sit "Layer"

Political Ideology

PartyID	Gender	Mean	N	Std. Deviation
Democrat	Female	3,8465	215	1,25664
	Male	3,7680	125	1,43198
	Total	3,8176	340	1,32226
Independent	Female	3,9527	169	1,24313
	Male	4,0417	144	1,30022
	Total	3,9936	313	1,26844
Republican	Female	4,4321	162	1,25543
	Male	4,6641	128	1,31183
	Total	4,5345	290	1,28359
Total	Female	4,0531	546	1,27464
	Male	4,1562	397	1,39291
	Total	4,0965	943	1,32597

Mange middelværdier i spil

- I eksemplet er der $2 \cdot 3 = 6$ celler i spil, med hver deres middelværdi:

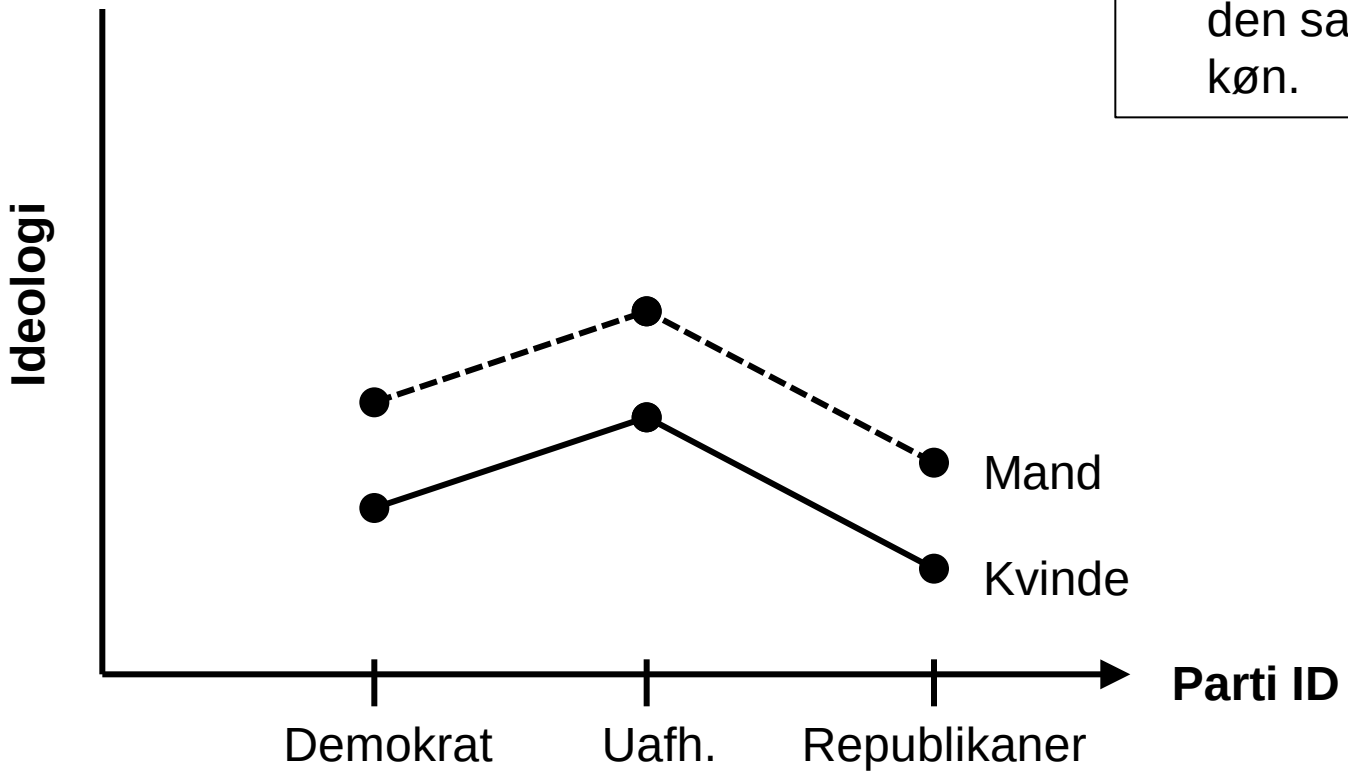
	Party ID		
Gender	Democrat	Independent	Republican
Female	μ_{FD}	μ_{FI}	μ_{FR}
Male	μ_{MD}	μ_{MI}	μ_{MD}

- En to-sidet variansanalyse handler om at undersøge, hvordan de to forklarende variable (Party ID og Gender) påvirker disse middelværdier.
- Der er **to slags effekter**:
  - **Hovedeffekter**: Isoleret effekten af en forklarende variabel
  - **Vekselvirkningseffekt**: Effekten af en variabel påvirkes af en anden variabel.

ANOVA model uden vekselvirkning

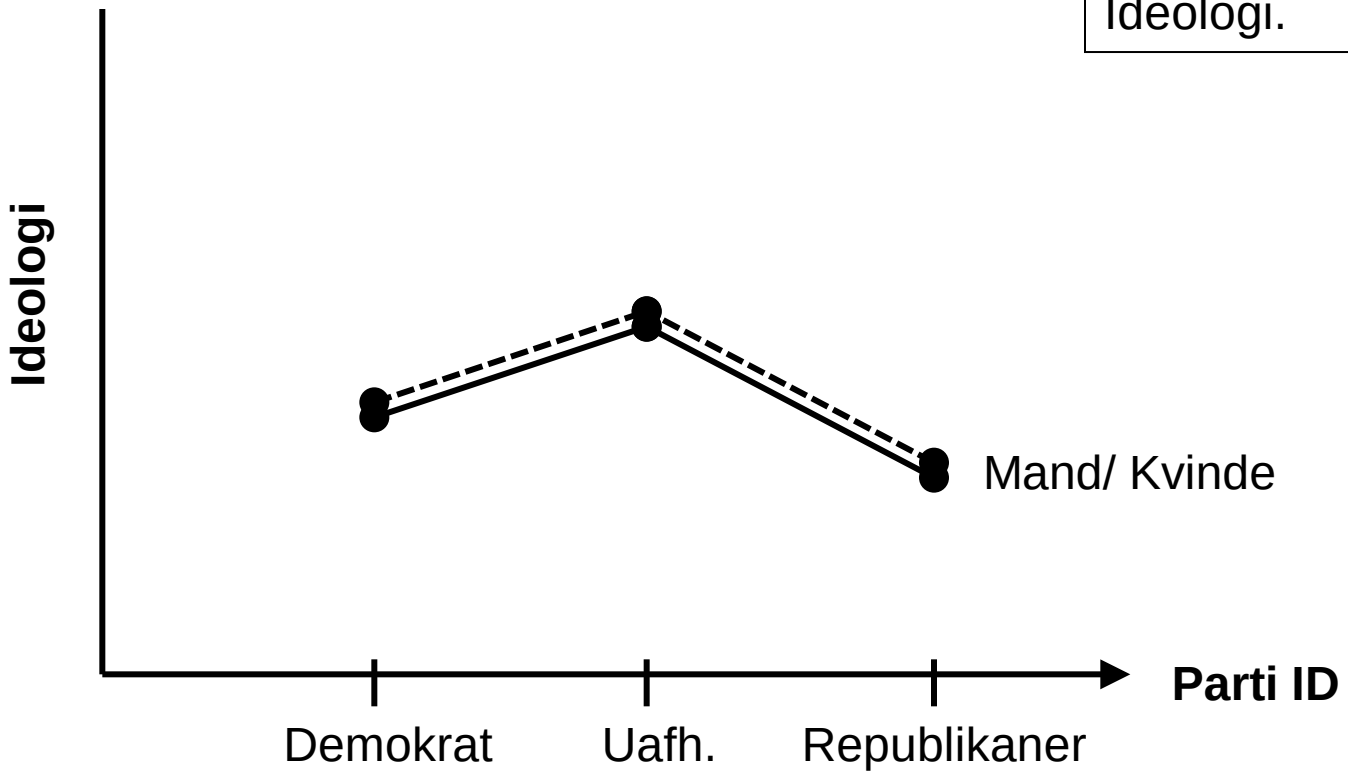
Fortolkninger:

1. Effekten af køn er den samme for alle Parti ID
2. Effekten af Parti ID er den samme for begge køn.



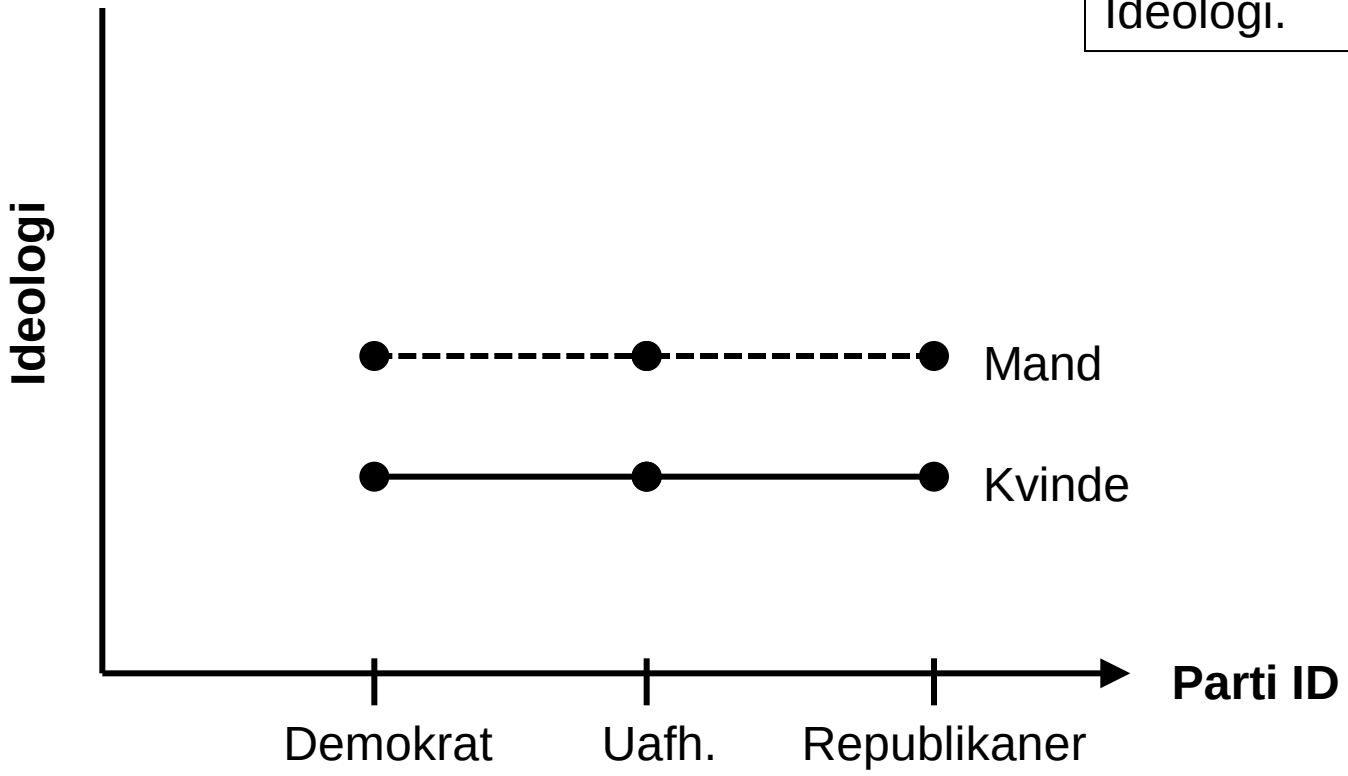
ANOVA kun med hovedeffekt A

Fortolkning:
Kun hovedeffekt A (Parti ID) har en betydning for Ideologi.



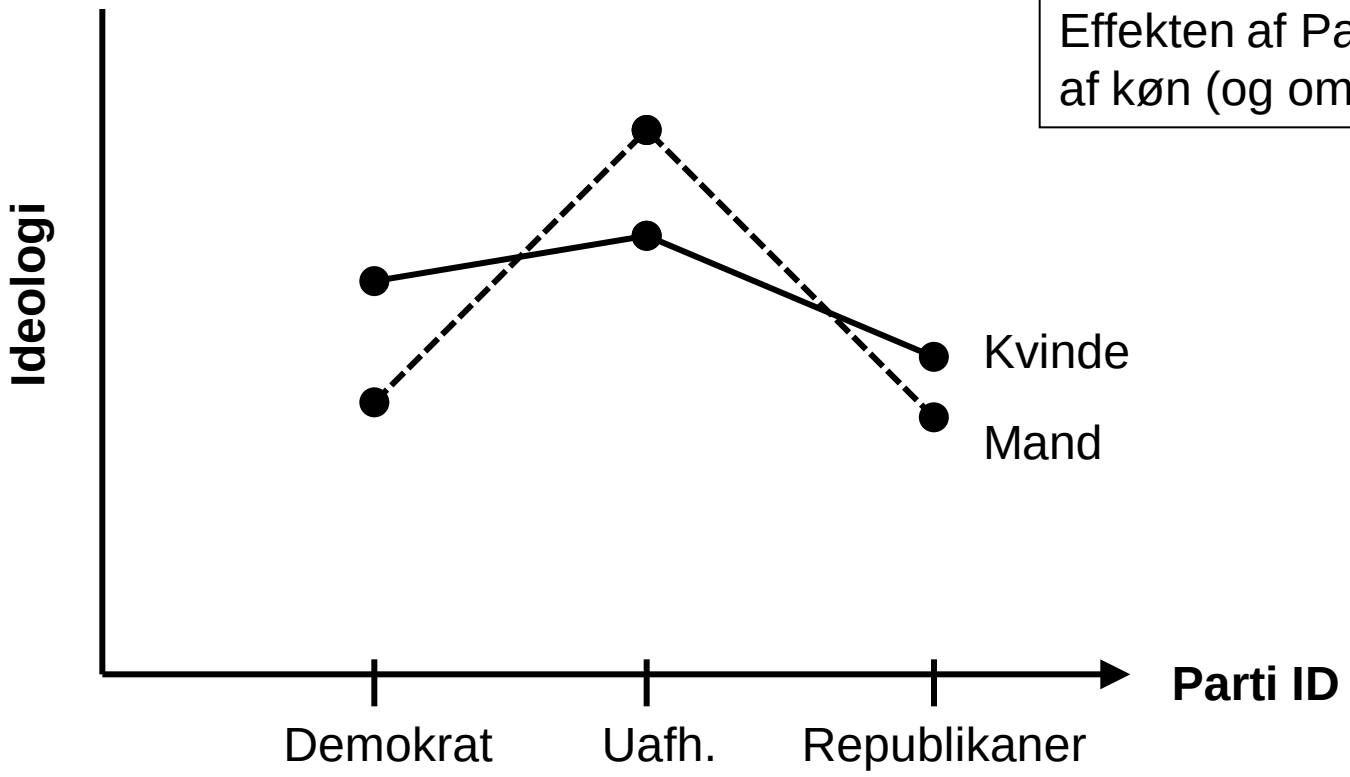
ANOVA kun med hovedeffekt B

Fortolkning:
Kun hovedeffekt B (Køn) har en betydning for Ideologi.

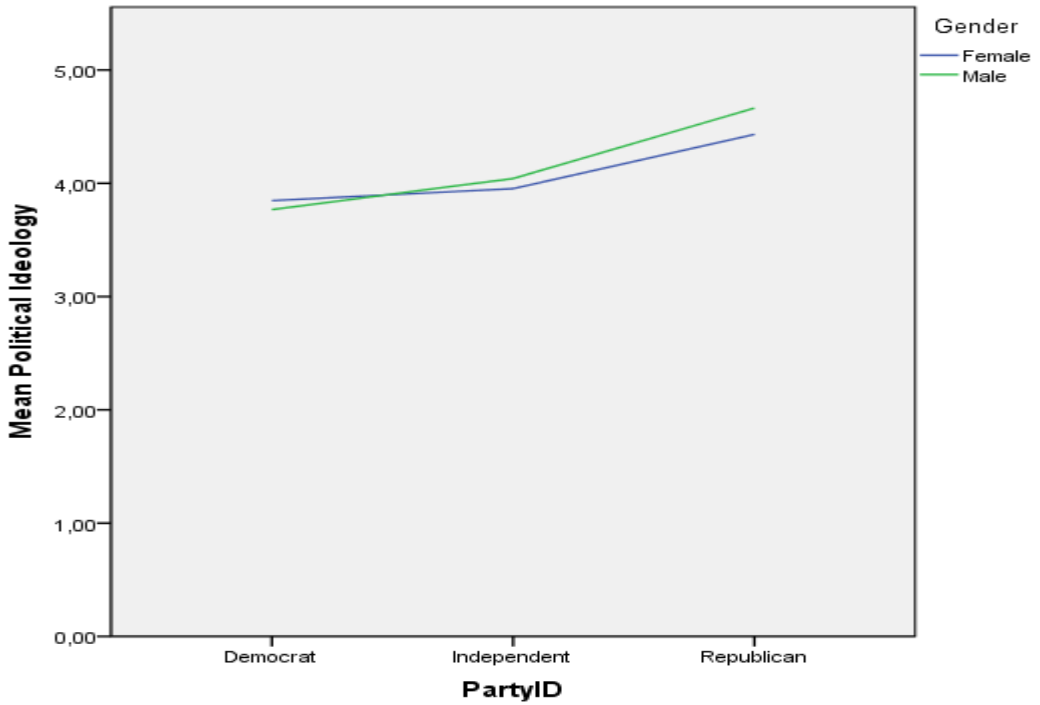


ANOVA model med vekselvirkning

Fortolkning:
Effekten af Parti ID afhænger af køn (og omvendt)



For data ser det sådan ud



- Ikke meget tegn på vekselvirkning
- Ikke meget tegn på effekt af køn
- En svag effekt af Party ID

Hypoteser og Antagelser

- **Antagelser**:
  - Observationerne i hver celle er **normalfordelte**
  - **Standardafvigelsen er konstant** på tværs af celler

- Vi tester **hypoteser** på formen
  - H_0: Ingen effekt af prediktor (=forklarene variabel)
  - H_a: Der *er* en effekt af prediktor

- Generelt: Antag vi har to prediktore, *A* og *B*: Vi vil teste
  - **Hovedeffekten** af prediktor ***A***
  - **Hovedeffekten** af prediktor ***B***
  - **Vekselvirkningseffekten** ml. ***A* og *B***.

Analyse-Strategi

- Slagplanen minder om den for multipel lineær regression:
- **Først** tester vi effekten af **vekselvirkningen**.
- Er vekselvirkningen signifikant, så tester vi ikke mere. Det giver ikke mening at teste hovedeffekter, hvis der er en vekselvirkning.
- Er vekselvirkningen *ikke* signifikant, så fjerner vi den fra modellen og tester de to tilbageværende **hovedeffekter**.

Hypoteser og Antagelser

- Vi tester altså **hypoteser** på formen
  - H_0: Ingen effekt af prediktor
  - H_a: Der *er* en effekt af prediktor

- **Teststørrelsen** er generelt på formen

$$F = \frac{\text{Mean square for prediktor}}{\text{Mean square error}}$$

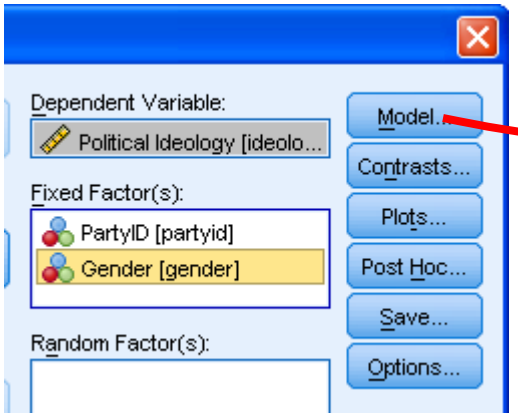
- Generelt gælder der at

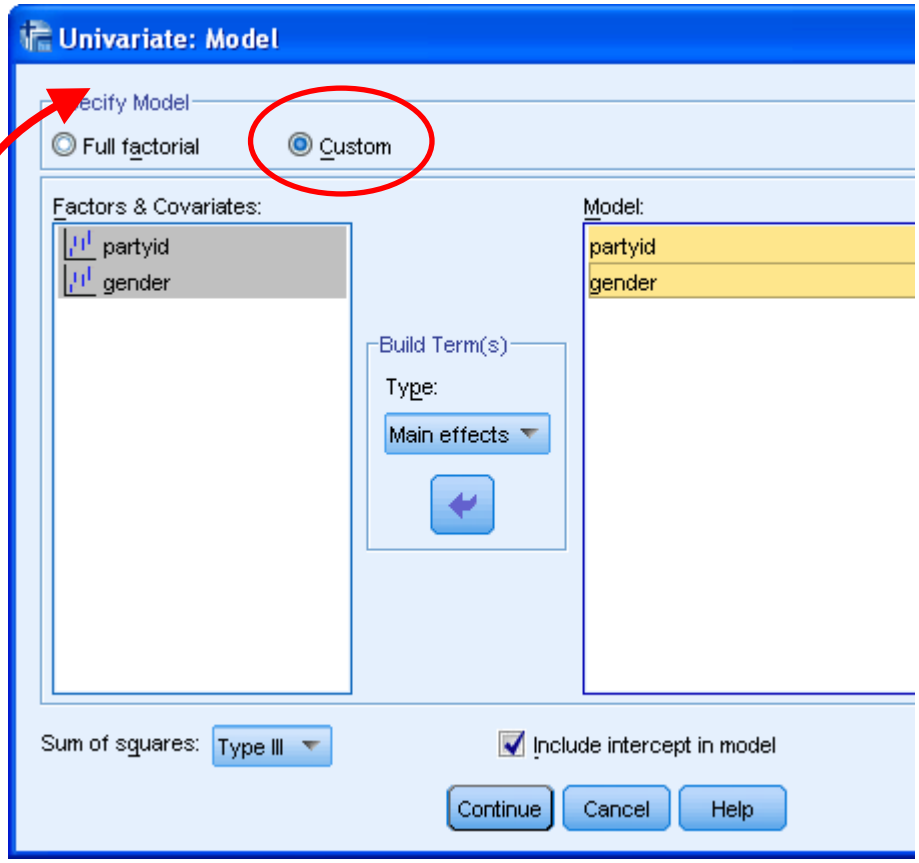
$$\text{Mean square} = \text{Sum of squares}/\text{df}$$

- SPSS finder Sum of Squares og antal frihedsgrader (*df*).

Eksempel: Model uden Vekselvirkning

- I SPSS er vekselvirkning taget med pr. default, så det skal der gøres noget ved.





- Vælg 'Custom' model.
- Vælg 'Main effects'
- Overfør de to 'factors'

SPSS: Resultat

Tests of Between-Subjects Effects

Dependent Variable:Political Ideology

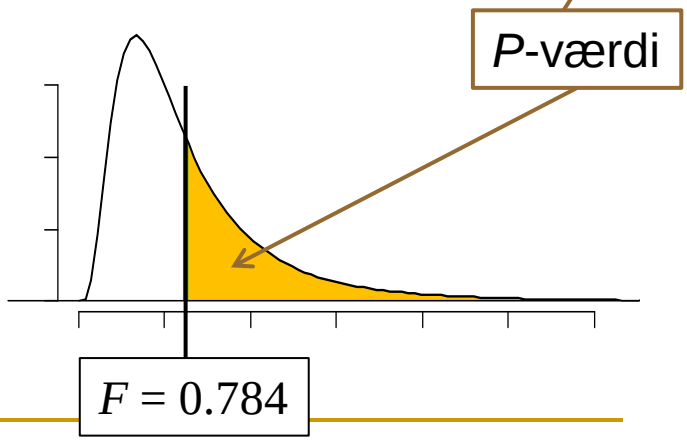
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	86,693[a]	3	28,898	17,289	,000
Intercept	15568,482	1	15568,482	9314,155	,000
partyid	84,252	2	42,126	25,203	,000
gender	1,311	1	1,311	,784	,376
Error	1569,525	939	1,671		
Total	17481,000	943			
Corrected Total	1656,218	942			

a. R Squared = ,052 (Adjusted R Squared = ,049)

- H_0: Ingen effekt af køn vs H_a: Der *er* en effekt af køn
- **Teststørrelse**

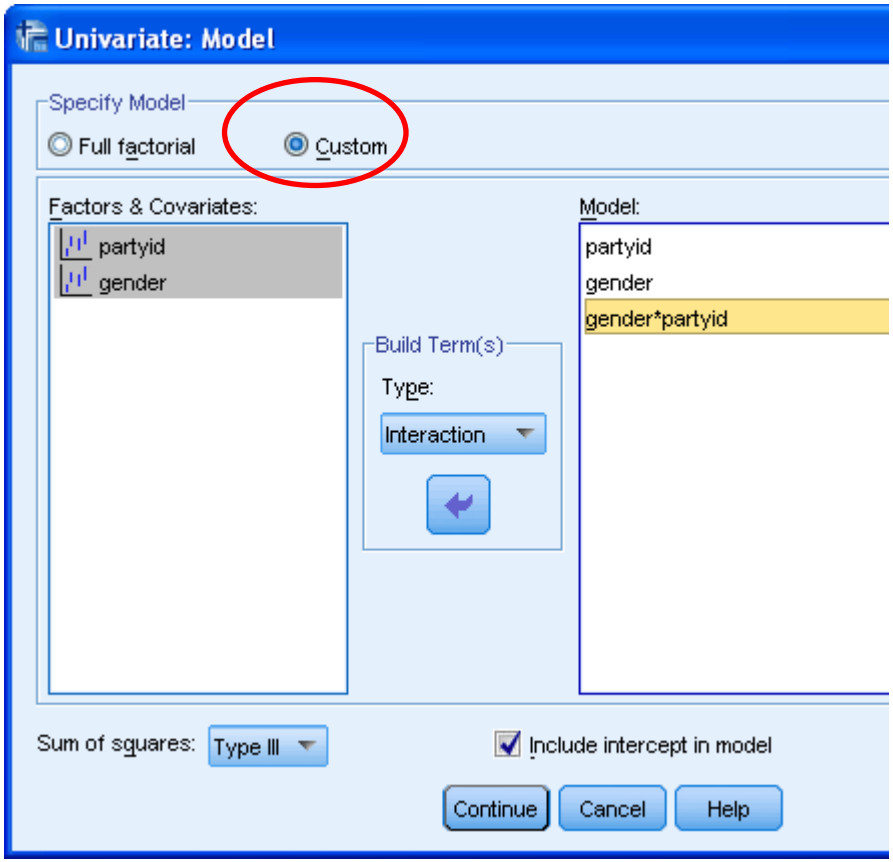
$$F = \frac{1.311}{1.671} = 0.784$$

- **Konklusion**: Da *P*-værdien > 0.05 kan vi *ikke* afvise H_0. Igen effekt af køn.



Test af vekselvirkning

- Vi spoler lige et trin tilbage.
- Antag at vi også inkluderer vekselvirkning i modellen:
- Enten skal man sikre sig at 'Full factorial' er valgt:
- Alternativt kan man selv angive modellen med vekselvirkning:
- Marker *både* **partyid** og **gender**, vælg **Interaction** og før over.



- **Vigtigt**: Det er vigtig at man *først* overfører hovedeffekterne og *derefter* vekselvirkningseffketer:

SPSS: Resultat

Tests of Between-Subjects Effects

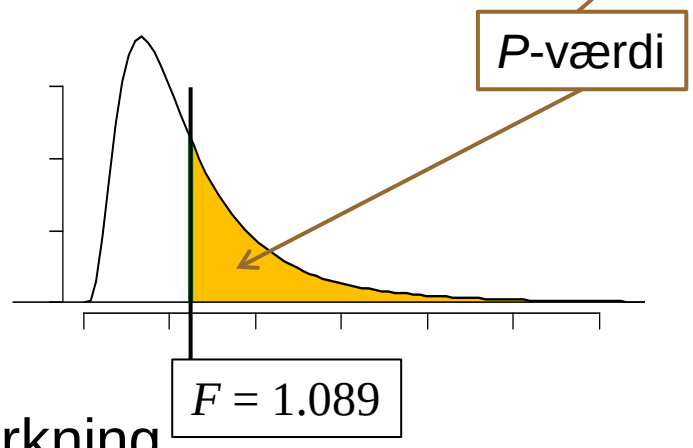
Dependent Variable:Political Ideology

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	90,332[a]	5	18,066	10,811	,000
Intercept	15452,314	1	15452,314	9246,406	,000
partyid	87,795	2	43,898	26,268	,000
gender	1,488	1	1,488	,891	,346
partyid * gender	3,640	2	1,820	1,089	,337
Error	1565,886	937	1,671		
Total	17481,000	943			
Corrected Total	1656,218	942			

a. R Squared = ,055 (Adjusted R Squared = ,049)

- H_0: Ingen effekt af vekselvirkning
- **Teststørrelse**

$$F = \frac{1.820}{1.671} = 1.089$$



- **Konklusion**: Da *P*-værdien > 0.05 kan vi *ikke* afvise H_0. Igen effekt af vekselvirkning.

To-sidet variansanalyse og Regression

- Først skal vi definere to sæt dummy-variable:
  - For Parti ID har vi to: p_1 og p_2
  - For Køn har vi en: s

Party ID	p_1 =	p_2 =
Democrat	1	0
Independent	0	1
Republican	0	0

Gender	s =
Female	1
Male	0

- To-sidet variansanalysemodel *uden* vekselvirkning:

$$E[y] = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s$$

Fortolkning

- Fortolkning af modellen:

$$E[y] = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s$$

- Tabel over middelværdier ifølge modellen:

E[y]=…	Demokrat p_1=1 p_2=0	Uafh. p_1=0 p_2=1	Republikaner p_1=0 p_2=0
Kvinde s = 1	$\alpha+\beta_1+\beta_3$	$\alpha+\beta_2+\beta_3$	$\alpha+\beta_3$
Mand s = 0	$\alpha+\beta_1$	$\alpha+\beta_2$	α

- **Bemærk:**
  - β_1 og β_2 angiver effekten af at være hhv. Demokrat og Uafh. i forhold til at være Republikaner (referencen).
  - Effekten af Parti ID den samme for begge køn.
  - β_3 angiver effekt af Kvinde i forhold til Mand.

Estimation

- Fra SPSS får vi:

Parameter Estimates

Dependent Variable:Political Ideology

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4,577	,090	51,016	,000	4,401	4,753
[partyid=1,00]	-,711	,104	-6,870	,000	-,914	-,508
[partyid=2,00]	-,542	,105	-5,146	,000	-,749	-,335
[partyid=3,00]	0[a]	.	.	.	.	.
[gender=1,00]	-,076	,086	-,886	,376	-,244	,092
[gender=2,00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

- **Estimerede model:**

$$\hat{y} = 4.577 - 0.711 \cdot p_1 - 0.542 \cdot p_2 - 0.076 \cdot s$$

- Effekten af at være Demokrat eller Uafh. i forhold til at være Republikaner er negativ.
- Effekten af Kvinde er negativ (i forhold til Mand).

Model med vekselvirkning

- To-sidet variansanalyse med vekselvirkning:

$$E[y] = \alpha + \beta_1 p_1 + \beta_2 p_2 + \beta_3 s + \beta_4 z_1 s + \beta_5 z_2 s$$

- Som i multipel lineær regression er vekselvirkning opnået ved at gange de to variable sammen.

E[y]=…	Demokrat p_1=1 p_2=0	Uafh. p_1=0 p_2=1	Republikaner p_1=0 p_2=0
Kvinde s = 1	$\alpha+\beta_1+\beta_3+\beta_4$	$\alpha+\beta_2+\beta_3+\beta_5$	$\alpha+\beta_3$
Mand s = 0	$\alpha+\beta_1$	$\alpha+\beta_2$	α

- Bemærk at vi har **6 parametre** og **6 celler**.
- Det er muligt med denne model **frit at tildele** hver celle en **middelværdi** uafhængigt af de andre celler.
- Man kalder sådan en model **mættet** – det er ikke muligt at gøre den mere kompliceret.

Estimation

Parameter Estimates

Dependent Variable:Political Ideology

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	4,664	,114	40,819	,000	4,440	4,888
[partyid=1,00]	-,896	,163	-5,512	,000	-1,215	-,577
[partyid=2,00]	-,622	,157	-3,963	,000	-,931	-,314
[partyid=3,00]	0[a]	.	.	.	.	.
[gender=1,00]	-,232	,153	-1,517	,130	-,532	,068
[gender=2,00]	0[a]	.	.	.	.	.
[partyid=1,00] * [gender=1,00]	,310	,211	1,472	,141	-,104	,725
[partyid=1,00] * [gender=2,00]	0[a]	.	.	.	.	.
[partyid=2,00] * [gender=1,00]	,143	,212	,675	,500	-,273	,559
[partyid=2,00] * [gender=2,00]	0[a]	.	.	.	.	.
[partyid=3,00] * [gender=1,00]	0[a]	.	.	.	.	.
[partyid=3,00] * [gender=2,00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

- Den estimerede model:

$$\hat{y} = 4.664 - 0.896 \cdot p_1 - 0.622 \cdot p_2 - 0.232 \cdot s + 0.310 \cdot p_1 \cdot s + 0.143 \cdot p_2 \cdot s$$