Anvendt Statistik
Lektion 8

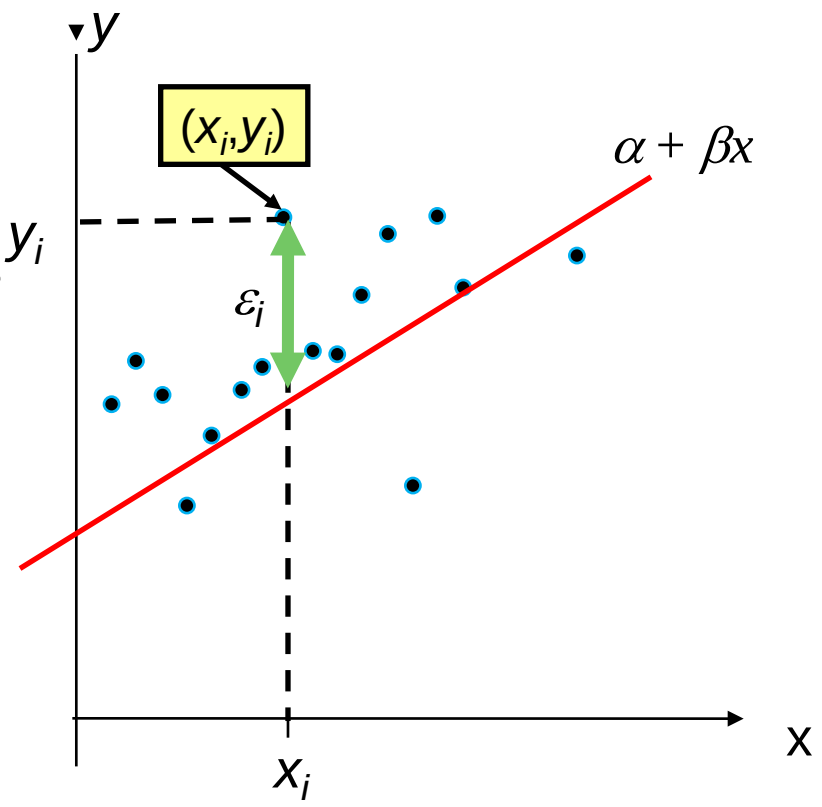
Multipel Lineær Regression

Simpel Lineær Regression (SLR)

- Sammenhængen mellem den afhængige variabel (y) og den forklarende variabel (x) beskrives vha. en SLR: ligger *ikke* præcist på regressionslinjen.
- **Regressionsmodel**:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

- *Fejlleddet* ε_i angiver afvigelsen mellem punktet (x_i, y_i) og linjen.



- Fejlledene er uafhængige og normalfordelte med middelværdi nul og standardafvigelse σ.

Multipel Lineær Regression (MLR)

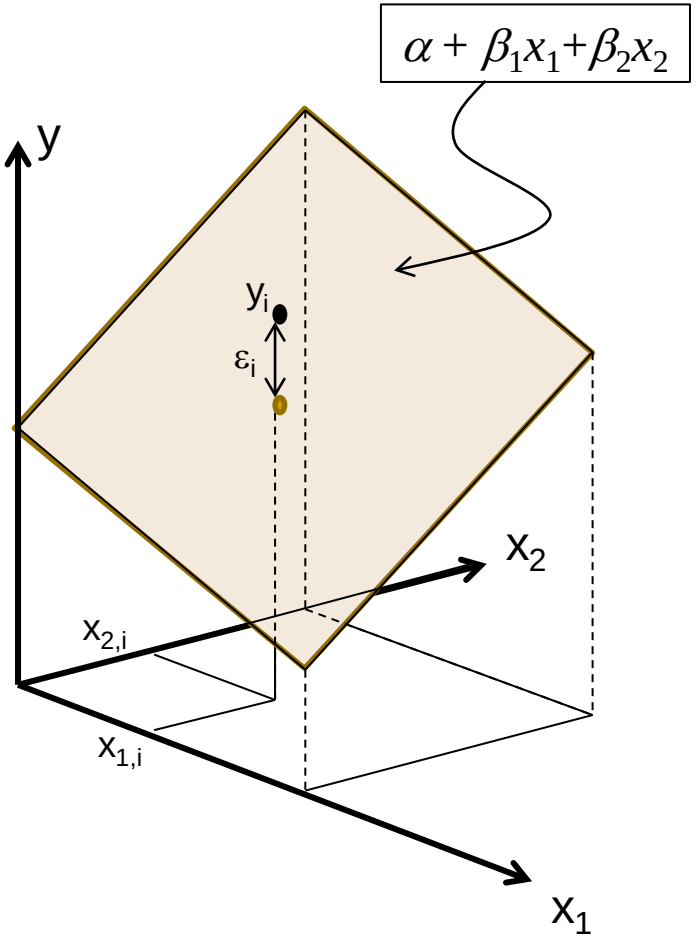
- Antag vi har
  - y : afhængig variabel
  - x_1 : første forklarende var.
  - x_2 : anden forklarende var.
- MLR model:

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

- Her:
  - $x_{1,i}$ er værdien af x_1 for *i*'te "person".
- Forventede værdi:

$$E[y] = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- Dvs. regressionsplanet angiver gennemsnittet for responsen



Fortolkning af β_j

- Antag vi har *k* forklarende variable:

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i$$

- Fortolkningen af β_j:
  - Hvis f.eks. x_1 øges med 1, så øges den forventede værdi af *y* med β_1, hvis $x_2, x_3, \ldots, x_k$ forbliver uændrede (og fortolkningen er naturligvis tilsvarende for $\beta_2, \beta_3, \ldots, \beta_k$).

Prædiktion og Residual

- **MLR model:**

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_k x_{k,i} + \varepsilon_i$$

- **Prædiktionsligningen er**

$$\hat{y}_i = a + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_k x_{k,i}$$

- Dvs. $\hat{y}_i$ er et estimat af $\mathrm{E}[y_i]$.

- **Residual:** $e_i = y_i - \hat{y}_i$
- Dvs. residualet er et estimat af ε_i .

Mindste kvadraters metode

- Definer summen af de kvadrerede residualer

$$SSE = \sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$$

- UK: Sum of Squared Errors
- SPSS: Sum of Squared Residuals

- Mindste kvadraters metode:
  - Vi vælger a, b_1, b_2, …, b_k, så SSE er mindst mulig.
  - Bemærk at

$$SSE = \sum_i \left(y_i - \left(a + b_1 x_{1,i} + b_2 x_{2,i} + \cdots + b_k x_{k,i}\right)\right)^2$$

Eksempel: Kriminalitet i Florida

- Tre variable
  - y : crime rate (kriminalitetsrate)
  - x1: education (uddannelse)
  - x2: urbanization (urbanisering)

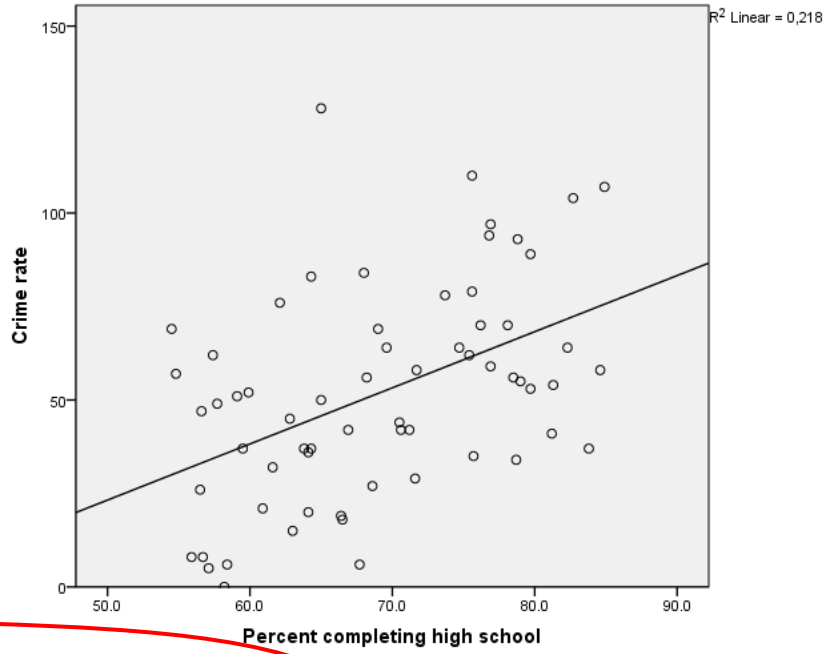
- I første omgang: Kriminalitetsrate og uddannelse

Eksempel: Kriminalitet i Florida (fortsat)

- En simpel lineær regression af kriminalitetsrate (y) mod uddannelse (x):
- Prædiktionsligning

$$\hat{y} = -51.8 + 1.50 \cdot x$$

- Dvs. jo mere uddannelse, jo mere kriminalitet…
- Effekten er statistisk signifikant.



Coefficients[a]

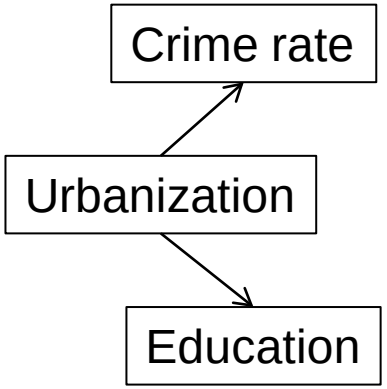
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-51,802	25,145		-2,060	,044
	Percent completing high school	1,501	,361	,467	4,156	,000

a. Dependent Variable: Crime rate

Eksempel: Kriminalitet i Florida (fortsat)

- **Teori**: Jo mere urbaniseret, jo mere kriminalitet og jo flere med lang uddannelse.
- Multipel lineær regression af kriminalitetsrate (y) mod både uddannelse (x_1) og urbanisering (x_2).
- Prædiktionsligning:

$$\hat{y} = 56.8 - 0.54 \cdot x_1 + 0.673 \cdot x_2$$



- Bemærk at effekten af uddannelse nu er negativ og ikke længere er signifikant (*P*-værdi >> 5%).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	56,801	29,400		1,932	,058
	Percent completing high school	-,541	,491	-,168	-1,103	,274
	Percent urban	,673	,128	,805	5,279	,000

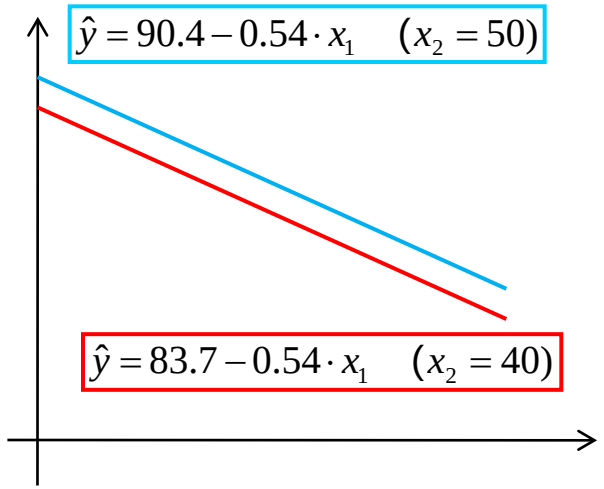
a. Dependent Variable: Crime rate

Eksempel: Kriminalitet i Florida (fortsat)

- Prædiktionsligning:

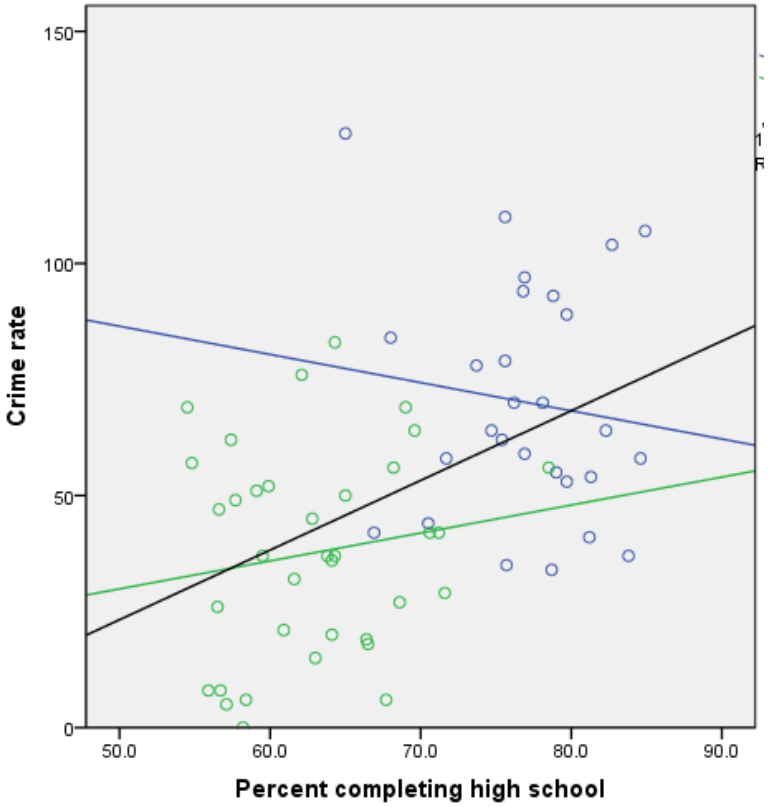
$$\hat{y} = 56.8 - 0.54 \cdot x_1 + 0.673 \cdot x_2$$

- Effekten af x_1 (uddannelse) er den samme for alle værdier af x_2 (ubanisering).
- For hver ekstra procent-point uddannede falder kriminalitetsraten med 0.54.
- Bemærk at effekten af x_1 (uddannelse) ændrede sig markant, da vi tilføjede x_2 (urbanisering). Det tyder på at der er en stærk sammenhæng mellem x_1 og x_2.



Simpsons paradoks - igen

- Sammenhæng mellem crime rate og uddannelse
- **Sort linje:**
  - SLR for alle data
- **Blå linje:**
  - SLR kun for områder med høj grad af urbanisering.
- **Grøn linje:**
  - SLR kun for områder med lav urbanisering.
- Bemærk hvor forskellig sammenhængen er i de to grupper.



Eksempel: Mentalt helbred

- Vi har tre variable:
  - y : Mental impairment (funktionsnedsættelse), afhængig var.
  - x_1 : Life events, første forklarende variabel.
  - x_2 : Socioøkonomisk status (SES), anden forklarende var.

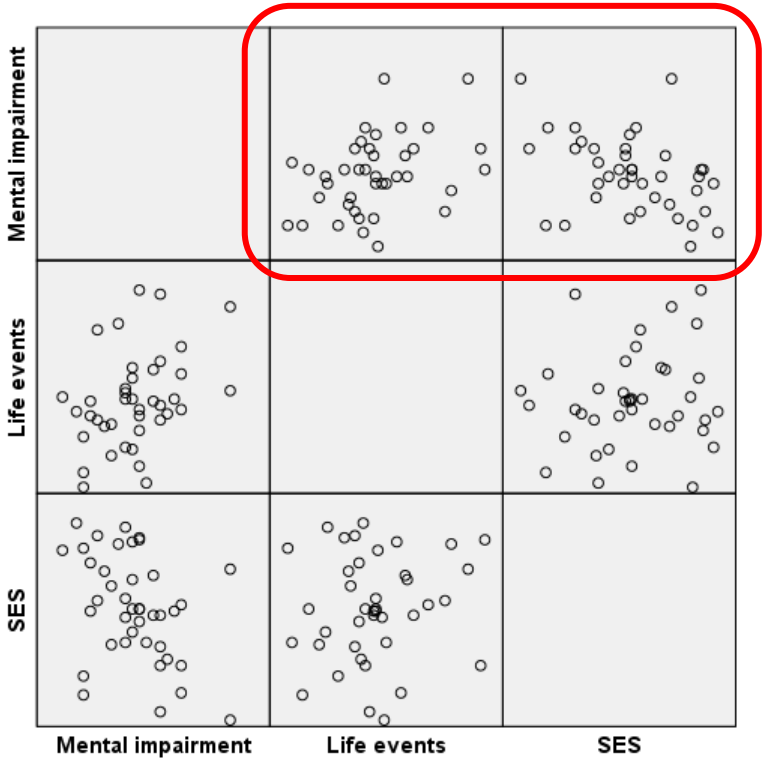
- Multipel lineær regressionsmodel:

$$y_i = \alpha + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \varepsilon_i$$

- MLR antager en lineær sammenhæng mellem y og hvert x_j.
- Vi starter med et scatter plot for hvert par af variable.

Scatterplot Matrix

- Graphs → Chart builder → Scatter/Dot →Scatterplot Matrix
- Ingen åbenlyse ikke-lineære sammenhænge.
- Ingen åbenbare sammenhænge i det hele taget…



- Problem: Plot viser sammenhængen mellem y og fx. x_1, hvor vi ignorer *v*ærdien af x_2.
- Vi har set, at vi ikke kan ignorere effekten af x_2, når vi ser på sammenhængen mellem *y* og x_1.

Partielt plot

- Estimeret model (eksempel med tre forklarende variable)

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + e$$

- Estimeret del-model (uden x_1)

$$y = \tilde{a} + \tilde{b}_2 x_2 + \tilde{b}_3 x_3 + \tilde{e}$$

- Regression af x_1 mod x_2 og x_3 (hvordan afhænger x_1 af x_2 og x_3)

$$x_1 = a^* + b_2^* x_2 + b_3^* x_3 + e^*$$

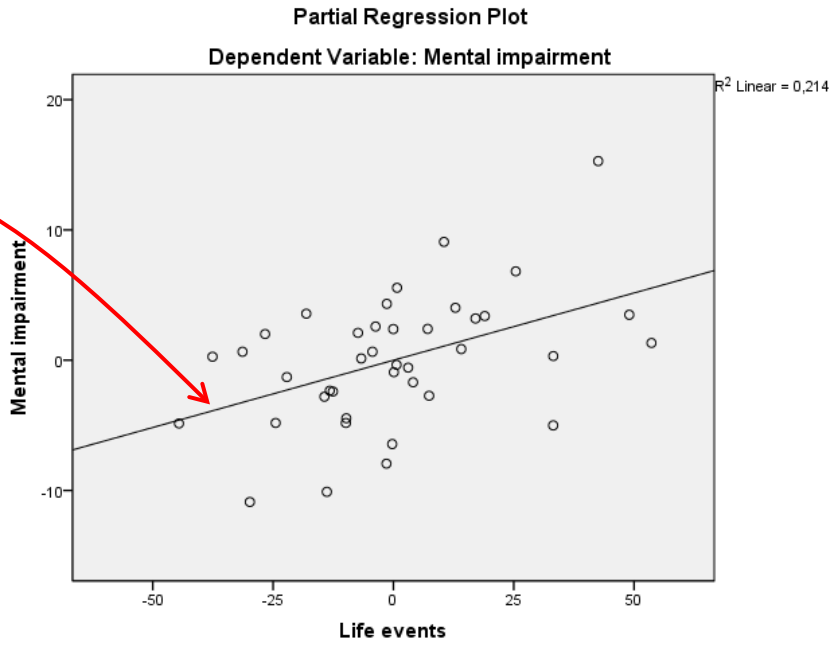
- Vi har to sæt residualer: $\tilde{e}$ (for y) og e^* (for x_1).
- **Ide**: plot $\tilde{e}$ mod e^*.

Partielt plot (fortsat)

- SPSS: Analyze → Regression → Linear → Plots → Produce all partial plots.
- Regression af $\tilde{e}$ mod e^* giver:

$$\tilde{e} = \hat{a} + \hat{b}e^* + \hat{e}$$

- **Interessant**: $\hat{b} = b_1$ Dvs. at hældningen i det partielle plot er den samme som effekten i den fulde model!



- **Bonus**: Check at residualerne varierer usystematisk og at variationen er den samme langs linjen.

SPSS output

- Simpel model – kun en forklarende variabel

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	23,309	1,807		12,901	,000
	Life events	,090	,036	,372	2,472	,018

a. Dependent Variable: Mental impairment

- Model med to forklarende variable:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28,230	2,174		12,984	,000
	Life events	,103	,032	,428	3,177	,003
	SES	-,097	,029	-,451	-3,351	,002

a. Dependent Variable: Mental impairment

Multipel korrelation

- Husk: Korrelation angiver hvor lineært afhængig to variable er.

- **Multipel korrelation** R for en lineær regression er korrelationen mellem de observerede y og de prædikterede $\hat{y}$.

- Bemærk: Den multiple korrelation kan ikke være negativ.

Multipel determinationskoefficient

- Den **totale variation** i *y*'erne:

$$TSS = \sum_i \left(y_i - \bar{y}\right)^2$$ (Total Sum of Squares)

- Den **uforklarede** del af variationen i *y*'erne:

$$SSE = \sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i e_i^2$$ (Sum of Squared Errors)

- Den **forklarede** del af variationen i *y*'erne:

$$TTS - SSE$$

- Multipel determinationskoefficient

$$R^2 = \frac{TSS - SSE}{TSS}$$

- **Fortolkning**: Andelen af den totale variation, der er forklaret.

Eksempel på R og R^2

- Lille model

$$y = \alpha + \beta_1 x_1 + \varepsilon$$

- $R^2 = 0.139$
- Dvs. 13.9% af variationen i mental impairment er forklaret af Life events.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,372[a]	,139	,116	5,133

a. Predictors: (Constant), Life events
b. Dependent Variable: Mental impairment

- Stor model

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

- $R^2 = 0.339$
- Dvs. 33.9% af variationen i mental impairment er forklaret af Life events *og* SES.
- Bemærk at R^2 er øget – vi kan forklare mere med flere variable.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,582[a]	,339	,303	4,556

a. Predictors: (Constant), SES, Life events
b. Dependent Variable: Mental impairment

Egenskaber for R og R^2

- R^2 er mellem 0 og 1
- Jo højere R^2, jo bedre kan modellen prædiktere y.
- $R^2 = 1$ betyder at $\hat{y} = y$ og alle residualer er nul.
- $R^2 = 0$ betyder at $b_1 = b_2 = \ldots = b_k = 0$.
- Når en variabel tilføjes modellen kan R^2 *ikke* falde.

Hypotesetest for MLR: *F*-test

- MLR model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

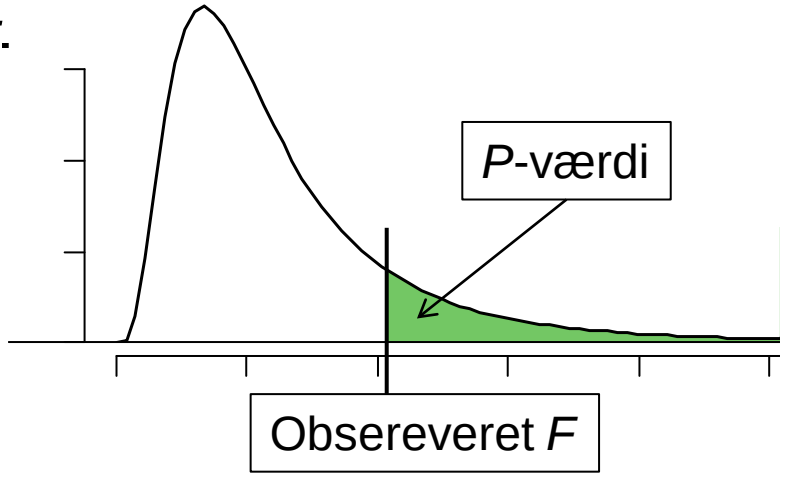
- Er der mindst en af x_j'erne der har en lineær sammenhæng med y?
- **Nul-hypotese**:
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

  y har ingen lineær sammenhæng med et eneste x_j.
- **Alternativ-hypotese**:
  - H_a: Mindst et $\beta_j \neq 0$

  y har en lineær sammenhæng med med mindst et af x_j'erne.
- **Teststørrelse**:
  - $$F = \frac{R^2 / k}{\left(1 - R^2\right) / \left(n - (k+1)\right)}$$

F-testet

- Hvis H_0 er sand, så følger *F* en *F*-fordeling.
- Som χ^2-fordelingen kan *F*-fordelingen kun tage positive værdier.
- Faconen på *F*-fordelingen er bestemt af to sæt frihedsgrader df_1 og df_2:
  - $df_1 = k$ = antal forklarende variable.
  - $df_2 = n - (k + 1) = n$ – antal parametre i modellen



F-test: Eksempel

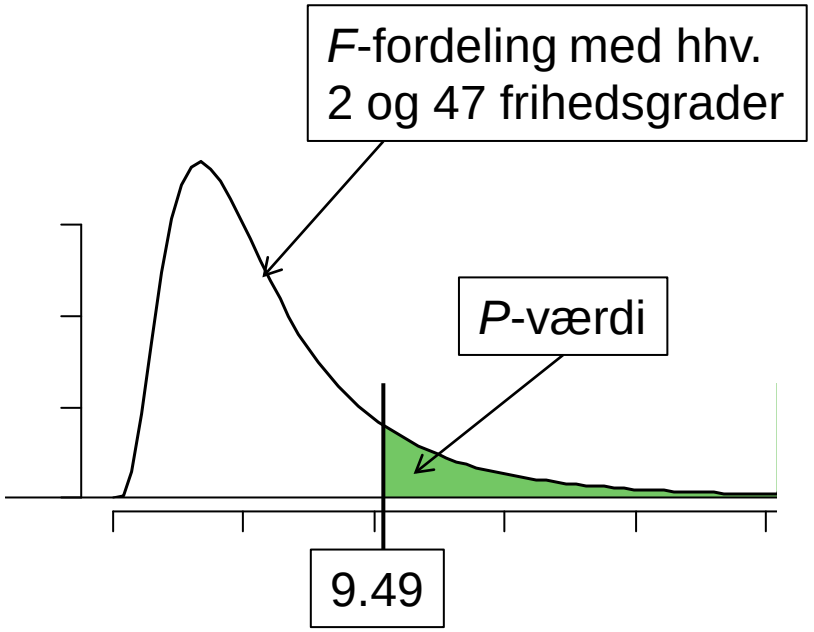
- **Model for mentalt helbred**:

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

- Fra SPSS har vi $R^2 = 0.339$
- Dvs.

$$F = \frac{R^2/k}{(1-R^2)/(n-(k+1))}$$

$$= \frac{0.339/2}{(1-0.339)/(40-3)} = 9.49$$



- *P*-værdien finder vi vha. SPSS (næste slide).
- Da *P*-værdien < 0.0005 afviser vi H_0, dvs. *y* har en lineær sammenhæng med mindst en af de to forklarende variable.

F-test i SPSS

- *F*-teststørrelsen kan omskrives:

$$F = \frac{R^2/k}{(1-R^2)/(n-(k+1))} = \frac{(TSS-SSE)/k}{SSE/(n-(k+1))}$$

$$= \frac{(1162.4-768.162)/2}{768.162/(40-3)} = 9.495$$

P-værdi

ANOVA[b]

SSE

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	394,238	2	197,119	9,495	,000[a]
	Residual	768,162	37	20,761		
	Total	1162,400	39			

TSS

a. Predictors: (Constant), SES, Life events
b. Dependent Variable: Mental impairment

Hypotesetest af βj

- MLR model:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- Er der en lineær sammenhæng mellem *y* og x_j?
- **Nul-hypotese**:
  - H_0: $\beta_j = 0$

y har ingen lineær sammenhæng med x_j.

- **Alternativ-hypotese**:
  - H_a: $\beta_j \neq 0$

y har en lineær sammenhæng med x_j.

- **Teststørrelse**:
  - $t = \dfrac{b_j}{se}$ ← Udregnes af SPSS
- Hvis H_0 er sand, så følger *t* en *t*-fordeling med *df* = *n*-(k+1)

Hypotesetest af βj : Eksempel

- **Model for mentalt helbred**:

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

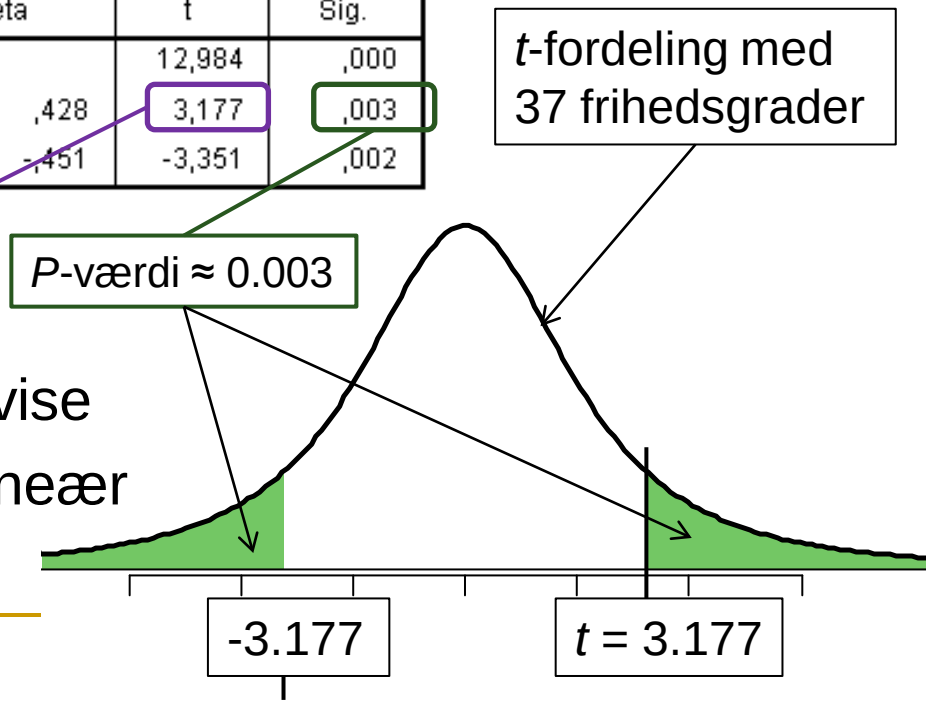
- Fra SPSS har vi b_1 = 0.103 og *se* = 0.032

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	28,230	2,174		12,984	,000
	Life events	,103	,032	,428	3,177	,003
	SES	-,097	,029	-,451	-3,351	,002

a. Dependent Variable: Mental impairment

- Dvs. t = 0.103/0.032 = 3.177
- Da *P*-værdien < 0.05, kan vi afvise H_0-hypotesen. Dvs. der *er* en lineær sammenhæng mellem y og x_1.



Estimation af σ

- Generelt er vores MLR model

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

- Vi antaget at fejlledene er normalfordelte med standardafvigelse σ.
- Et estimat af σ er

$$s = \sqrt{\frac{SSE}{n-(k+1)}}$$

- Eksempel:

$$s = \sqrt{\frac{768.162}{40-3}}$$

$$= \sqrt{20.761} = 4.56$$

ANOVA[b]

Model		Sum of Squares	df	Mean Square
1	Regression	394,238	2	197,119
	Residual	768,162	37	20,761
	Total	1162,400	39	

a. Predictors: (Constant), SES, Life events
b. Dependent Variable: Mental impairment

Vekselvirkning

- Der er **vekselvirkning** (også kaldet interaktion) mellem to forklarende variable, x_1 og x_2, for *y*, hvis effekten af x_1 på *y* ændrer sig når x_2 ændrer sig.

- Simpel **vekselvirkningsmodel**:

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

- Hvor kommer interaktionen ind i billedet?
- Omskriv modellen til

$$y = (\alpha + \beta_2 x_2) + (\beta_1 + \beta_3 x_2)\, x_1 + \varepsilon$$

- **Bemærk**: Hældningen er $\beta_1 + \beta_3 x_2$, dvs. effekten af x_1 på y ændrer sig, når x_2 ændres.

Vekselvirkning: Eksempel

- Simpel **vekselvirkningsmodel**:

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

- Vha. Transform → Compute variable skaber vi variablen
  `x1x2 = x1*x2`
- Følgende test viser at interaktionen ikke er signifikant:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	26,037	3,949		6,594	,000
	Life events	,156	,085	,646	1,826	,076
	SES	-,060	,063	-,280	-,965	,341
	X1X2	-,001	,001	-,307	-,668	,509

a. Dependent Variable: Mental impairment

- Da vekselvirkningen *ikke er* signifikant, kan man vælge at fjerne den.
- Hvis vekselvirkningen *er* signifikant, beholder vi det. Desuden giver det ikke mening at teste de enkelte led (x_1 og x_2).

Vekselvirkning: Eksempel (fortsat)

- Estimeret vekselvirkningsmodel:

$$y = 26.037 + 0.156 \cdot x_1 - 0.060 \cdot x_2 - 0.01 \cdot x_1 \cdot x_2$$

- Fortolkning:
- Når vi øger x_2, så
  - Reduceres skæringspunktet
  - Reduceres hældningen.