

Hvad kan der udledes: Opgave med lineær regression

Udled-opgaven indgår i fællesdelen af studentereksamensopgaverne, og besvarelsen skal ikke være specielt lang (cirka 350 ord), da den er én ud af to fællesdelsopgaver, der tilsammen maksimalt må fylde 700 ord.

?

HVAD ER EN UDLEDNINGSOPGAVE?

I denne opgavetype skal du udlede mønstre eller en udvikling ud fra en tabel eller figur. Disse tabeller eller figurer skal behandles forskelligt, afhængigt af om du bliver bedt om at arbejde med:

1. en tabel (med/uden beregninger)
2. en lineær regression
3. en beregning af statistisk usikkerhed (konfidensintervaller).

Fælles i alle tre typer udled-opgaver er, at du skal forklare de ting, du udleder, med anvendelse af faglig viden.

Eksempel på en lineær regressionsopgave:

Hvilken sammenhæng mellem dem, der mener, at der finder høj grad af diskrimination mod sorte, homoseksuelle og immigranter sted, og dem, der stemte på Donald Trump, kan der udledes af regressionen?

Din besvarelse skal understøttes af et diagram med en lineær regression, der viser sammenhængen mellem andelen, der mener, at der finder høj grad af diskrimination sted, og andelen, der stemte på Donald Trump ved præsidentvalget i 2016. Hvilken sammenhæng mellem andelen af vælgere, der mener, at der finder høj grad af diskrimination sted, og dem, der stemte på Donald Trump, kan der udledes af regressionen?

Du skal anvende viden om værdipolitik.

Når du får opgaven stillet, kan den være udformet på to måder. Enten får du udleveret en opgave med et diagram af en lineær regression, der allerede er lavet, som du skal arbejde med (du kan se, hvordan et diagram for tabel 3.1 ser ud, under 'det gode eksempel' på side 28).

Eller også får du udleveret en tabel med to ko-

Tabel 3.1:

Andel af amerikanere, der mener, at der finder diskrimination sted, og andel af stemmer på Donald Trump ved præsidentvalget i 2016 fordelt på udvalgte amerikanske stater (i procent).

Stater	Andel, der mener, at der finder høj grad af diskrimination sted mod sorte, homoseksuelle og immigranter	Andel, der stemte på Donald Trump ved præsidentvalget i 2016
Alabama	34	62,1
Arkansas	35	60,6
Californien	48	31,6
Connecticut	49	40,9
Georgia	44	51
Idaho	40	59,3
Illinois	44	38,8
Kansas	39	56,7
Louisiana	34	58,1
Massachusetts	46	32,8
New Jersey	44	41,4
North Dakota	27	63
Oregon	44	39,1
South Carolina	35	54,9
Vermont	38	30,3
Washington	41	36,3
Wyoming	28	68,2

Kilde: Carlson, Populisme og identitetspolitik, Columbus 2020.

lonner med tal i, du selv skal lave til et diagram, som du derefter skal bruge til selve udledningen (du kan se en konkret guide til at udarbejde et diagram med regression i kapitel 10).

Vi omtaler hver af talkolonnerne som en variabel. I det her tilfælde er den første variabel 'andel, der mener, at der finder høj grad af diskrimination sted mod sorte, homoseksuelle og immigranter', og den anden variabel er 'andel, der stemte på Donald Trump i 2016'.

Når vi laver en lineær regression, undersøger vi, om der er en statistisk sammenhæng mellem de to variable – det, vi også kalder samvariation eller korrelation. Det vil sige, at vi ser på, i hvor høj grad den ene variabel statistisk kan forklare den anden. Det gør vi ved at sætte den ene variabel som den uafhængige (den, der ikke er afhængig af den anden) og den anden variabel som den afhængige (altså den variabel, der afhænger af den anden).

Uafhængig variabel → afhængig variabel
(påvirker)

I eksemplet ovenfor undersøger vi således, om der er en sammenhæng mellem dem, der synes, at der finder diskrimination sted, og dem, der stemmer på Trump – du kan se resultatet af en sådan lineær regression under 'det gode eksempel' på side 28.

Det er vigtigt at bemærke her, at bare fordi vi kan se en statistisk sammenhæng mellem variablene (altså at den lineære regression viser, at de to variable 'følges ad'), så er det ikke nok til at kunne konkludere, at den ene variabel med sikkerhed medfølger den anden; altså at der er tale om kausalitet (for eksempel at holdningen til diskrimination er årsag til stemmeadfærd; om man stemmer på Trump). Der skal andre kriterier også opfyldes såsom tidsrækkefølge (at den uafhængige variabel går forud for den afhængige variabel i tid), tjek for spuriøse sammenhænge (hvor der nok er en statistisk sammenhæng mellem de to variable, men hvor den uafhængige variabel ikke er årsag til den afhængige) og teoretisk forklaring; alle sammen ting, du ikke skal komme nærmere ind på i din besvarelse, men som er gode at kende til.

Sådan gør du – trin for trin

Når du skal skrive en udled af lineær regression-opgave, kan du følge denne opskrift:

1. Læs opgaveformuleringen grundigt for at se følgende:

- * Hvilke variable optræder i tabellen? Det er altid en sammenhæng mellem disse to, du skal udlede noget ud fra
- * Hvilken faglig viden, du skal komme ind på (for eksempel viden om værdipolitik).

2. Beskriv sammenhængen mellem årsag og virkning

- * Normalt laves opgaverne til eksamen sådan, at den uafhængige variabel altid nævnes først i opgaveformuleringen
- * Den uafhængige variabel (x) betragtes altid som årsag til den afhængige variabel (y), således at $x \rightarrow y$
- * Den uafhængige variabel (x) er den vandrette, og den afhængige variabel (y) er den lodrette i dit diagram
- * I eksemplet ovenfor vil det være 'holdningen til, om diskrimination er et problem', der fører til, hvem man stemmer på. Under 'det gode eksempel' kan du se, hvordan en sådan opgave kan løses.

3. Udarbejd diagrammet med den lineære regression i Excel (hvis det ikke allerede er lavet)

- * Den uafhængige variabel (x) skal stå i den første kolonne til venstre, og den afhængige

variabel (y) skal stå i den anden kolonne til højre, når du indsætter tabellen i Excel

- * Husk at markere i Excel, sådan at forklaringsgrad og ligning kommer med i figuren
- * Du kan læse om, hvordan du konkret udarbejder diagrammet med den lineære regression, i kapitel 10.

4. Beskriv punkternes placering

- * Nu skal du sætte diagrammet med din lineære regression ind i dit dokument og starte med at beskrive punkterne
- * Ligger de tæt om tendenslinjen eller spredt omkring den? Ligger punkterne som et bredt bælte eller som en større bisværme? Samler de sig i nogle grupper, og hvad kan det skyldes?
- * Er der outliers (punkter, der ligger længe fra tendenslinjen end de øvrige punkter)?
- * Er der en positiv eller negativ tendens? (Er fortegnet på hældningskoefficienten positiv eller negativ?)

5. Tolk på R^2 -værdien

- * Nu skal du beskrive din tolkning af R^2 -værdien
- * R^2 -værdien er den lineære models forklaringsgrad; den siger noget om, hvor stor en del af variationen i den afhængige variabel der beskrives af tendenslinjen – med andre ord, hvor stærk er den statistiske sammenhæng mellem de to variable?
- * En R^2 -værdi på 1 viser en fuldstændig sammenhæng, mens en R^2 -værdi på 0 viser ingen sammenhæng overhovedet
- * Jo tættere punkterne ligger på tendenslinjen, jo mere vil R^2 -værdien nærme sig 1, og jo mere spredt punkterne ligger i forhold til tendenslinjen, jo mere vil R^2 nærme sig 0

- * Derfor det vigtigt at tjekke for outliers, fordi disse kan mindske R^2 -værdien, hvilket betyder, at uden den outlier, så ville sammenhængen mellem variablene være stærkere
- * Hvis R^2 -værdien er på for eksempel 0,40, så kan den lineære model forklare 40 procent af datapunkternes variation. Restvariationen på 60 procent kan modellen ikke forklare – de kan skyldes andre ukendte variable eller tilfældige variationer.

6. Tolk ligningen

- * Nu skal du beskrive din tolkning af ligningen
- * 'a' angiver hældningskoefficienten. Hældningskoefficienten er tendenslinjens hældning. Hældningskoefficienten viser, hvor meget y-værdien vokser eller falder, når du går en enhed (for eksempel procent) ud på x-aksen (se uddybning under 'det gode eksempel' på side 28)
- * 'b' angiver skæringen med y-aksen. Den skal du undlade at kommentere.

7. Anvend viden

- * Når du skal anvende faglig viden, så er det for at *forklare, hvorfor* der er eller ikke er en sammenhæng mellem de to variable
- * Til at forklare disse ting skal du anvende den faglige viden, du har lært om i timerne
- * Denne faglige viden kan være konkret viden om samfunds-faglige forhold, og det kan være samfunds-faglig teori
- * I opgaveformuleringen vil der altid stå, hvilken faglig viden du skal anvende. Det skal du se som en hjælp til dig, så du ikke selv skal finde ud af, hvilken viden du skal bruge.

Det gode eksempel: Udled af lineær regression



Opgaveformulering

På baggrund af tabel 3.2 skal du udarbejde et diagram med en lineær regression, der viser sammenhængen mellem andelen, der stemte på Donald Trump ved præsidentvalget i 2016, og andelen, der mener, at der finder høj grad af diskrimination sted mod sorte, homoseksuelle og immigranter, og andelen, der stemte på Donald Trump ved præsidentvalget i 2016.

Hvilken sammenhæng mellem dem, der mener, at der finder høj grad af diskrimination sted og dem, der stemte på Donald Trump, kan der udledes af regressionen?

Du skal anvende viden om værdipolitik.

Besvarelse

○ Sammenhængen mellem variablene

Figur 3.1 viser sammenhængen mellem andelen af amerikanere, der mener, at der finder diskrimination sted, fordelt på udvalgte amerikanske stater, og andelen af stemmer på Donald Trump ved præsidentvalget i 2016 opgjort i procent.

Andelen, der mener, der finder diskrimination sted, er sat som den uafhængige variabel, og andelen af stemmer på Donald Trump er sat som den afhængige variabel.

Det fremgår af figuren, at der er en negativ korrelation mellem de to variable, hvilket betyder, at jo større procentandel i en stat der mener, at der finder diskrimination sted, jo mindre andel af stemmer er gået til Donald Trump. Dette fremgår også af det negative fortegn på hældningen til regressionslinjen.

○ Beskrivelse af punkternes placering

Figurens datapunkter har en spredning omkring tendenslinjen som en cigar med en enkelt outlier (Vermont), der medvirker til en lavere forklaringsgrad. Modellen har en R^2 -værdi på 0,61, hvilket betyder, at 61 procent af variationen i den afhængige variabel kan forklares ved hjælp af den uafhængige variabel. Restvariationen på 39 procent kan modellen ikke forklare – de kan skyldes andre ukendte variable, eller tilfældige variationer.

○ Tolkning af R^2 -værdien

Det fremgår af regressionsligningen ($y = -1,52x + 108,39$), at for hver gang andelen i en stat, der mener, at der findes diskrimination sted, vokser med 1 procentpoint, så vil andelen af befolkningen, der stemmer på Donald Trump, falde med 1,52 procentpoint.

○ Tolkning af ligningen

○ Anvendelse af viden

En substantiel forklaring på den tendens, vi ser i figuren, kan lyde på, at vælgerens holdning til, hvorvidt der finder diskrimination sted, har betydning for, hvor vælgeren placerer sig på den værdipolitiske skala. Således, at hvis man anerkender diskrimination, vil man som oftest synes, det er et samfundsmæssigt problem, og så vil man placere sig i den moderne del af den værdipolitiske skala. Placeringen her er afgørende for, at man ikke stemmer på en kandidat som Donald Trump, da hans politik netop er kendetegnet ved et fokus på de traditionelle værdier.

BEDØMMELSE

Når din lærer retter din udled af lineær regression-opgave, vil hun eller han lægge vægt på følgende:

- * at du har udarbejdet et korrekt diagram med en lineær regression, ligning og R^2 (hvis diagrammet ikke allerede er lavet)
- * at placering af uafhængig og afhængig variabel er meningsfuld
- * at tolkningen af R^2 -værdien er korrekt
- * at tolkningen af ligningen er korrekt
- * at du anvender viden om det, du bliver bedt om
- * at din besvarelse er velstruktureret (brug gerne samme struktur som ovenfor i 'det gode eksempel').

TYPISKE FEJL VED UDLED AF LINEÆR REGRESSION

- * at du bytter om på uafhængig og afhængig variabel (x-aksen viser *altid* den uafhængige, der medfører den afhængige, som *altid* er vist på y-aksen)
- * at du ikke tolker R^2 -værdien korrekt
- * at du ikke tolker ligningen korrekt
- * at din gennemgang af den lineære regression ikke er velstruktureret.

!

Tjekliste

- Du har udarbejdet et diagram med en lineær regression (indeholdende tendenslinje, ligning og R^2 -værdi) og sat den ind i besvarelsen
- Diagrammet har en diagramtitel, aksetitler samt måleenheder
- Du har styr på, at x-aksen (den vandrette akse) er den uafhængige, og at y-aksen (den lodrette) er den afhængige
- Du beskriver, om sammenhængen mellem de to variable er positiv eller negativ – og hvad det betyder
- Du beskriver datapunkterne: Er de tæt på tendenslinjen? Er der outliers?
- Du tolker på R^2 -værdien
- Du tolker på ligningen
- Du giver en faglig/teoretisk forklaring på, hvorfor vi ser den sammenhæng, vi gør, i den lineære regression
- Din besvarelse holder sig nogenlunde inden for 350 ord.

Øvelser

3. Miniøvelser i udled af lineær regression-opgaven

- Kig på tabel 3.3 og overvej, hvilken variabel du vil sætte som den uafhængige og afhængige i en lineær regression – skattetryk eller årlige arbejdstimer? Begrund dit valg
- Lav et diagram med en lineær regression i Excel ud fra tabel 3.3. Se side XX i kapitel 10, hvordan du konkret gør dette
- Fortolk R^2 -værdien. Hvad siger den om sammenhængen mellem de to variable?
- Kig på din lineære regression – er der nogle outliers? Prøv at fjerne disse, og beskriv, hvad det gør ved R^2 -værdien. Forklar, hvorfor den ændrer sig.
- Fortolk ligningen. Forklar, hvad hældningskoefficienten fortæller os?
- Hvilken overordnet faglig viden/teori kan belyse tabellens indhold? Spørg for eksempel dig selv, hvilke lande der hænger sammen i klynger – og hvorfor.

Tabel 3.3:

Skattetryk i procent og gennemsnitlige arbejdstimer om året i udvalgte lande.

	Skattetryk	Gennemsnitlige årlige arbejdstimer
Frankrig	55,5	1472
Sverige	54,7	1621
Danmark	54,4	1410
Finland	52,0	1653
Østrig	51,1	1601
Slovenien	49,5	1682
Ungarn	48,3	1761
Italien	46,2	1730
Storbritannien	46,0	1676
Tyskland	45,0	1363
Polen	39,5	1928
Slovakiet	38,3	1740
Estland	37,2	1855
Irland	36,0	1879
Tjekkiet	40,9	1770
Portugal	41,2	1842

Note: Skattetryk er skatternes andel af BNP.

Kilde: Undervisningsministeriet. Samfundsfag A, 29. maj 2018 (revideret udgave).

Tabel 3.4:

Bruttoledige og Danmarks statsgæld i procent af BNP, 2007-2016.

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Bruttoledige	85	86	155	164	157	161	148	129	116	113
Danmarks statsgæld i % af BNP	27,3	33,4	40,4	42,9	46,4	45,2	44,0	44,0	39,5	37,7

Kilde: Undervisningsministeriet. Samfundslag A, Vejledende opgavesæt 1.

4. Konkrete opgaver

Når du skal besvare nedenstående opgaver, kan du gå tilbage og se på overblikket over udled af lineær regression-opgavens dele (se side 26-27) og på det gode eksempel (se side 28).

Opgave 1:

På baggrund af tabel 3.4 skal du udarbejde et diagram med en lineær regression, der viser sammenhængen mellem antal ledige og statsgælden i Danmark.

Hvilken sammenhæng mellem antal ledige og statsgælden i Danmark kan der udledes af regressionen?

Du skal anvende viden om økonomiske sammenhænge.

Opgave 2:

Hvad kan der på baggrund af tabel 3.5 udledes om sammenhængen mellem prisen på cigaretter og andel rygere i udvalgte lande?

Besvarelsen skal understøttes af et diagram med en lineær regression, der viser sammenhængen mellem prisen på cigaretter og andel rygere i udvalgte lande.

Du skal i besvarelsen anvende viden om markedsmekanismen og politisk påvirkning heraf.

Tabel 3.5:

Pris pr. pakke cigaretter (danske kroner) og andel rygere (i procent) i udvalgte lande.

	Pris pr. pakke cigaretter (danske kroner)	Andel rygere (i procent)
Norge	90	12,5
Sverige	37	8,7
Danmark	40	12,3
Storbritannien	70	13,7
Frankrig	45	20,5
Tyskland	37	15,0
Italien	30	17,4
Grækenland	22	27,0
Spanien	30	22,2
Holland	45	17,2
Belgien	37	16,8
Letland	14	24,1
Ungarn	22	25,8
Cypern	30	25,2
Rumænien	22	19,8
Østrig	30	23,9

Note: I den angivne pris indgår landets afgifter på cigaretter, og priserne er købekraftskorrigerede.

Kilde: Undervisningsministeriet. Samfundslag A, Vejledende opgavesæt 2 (uddrag).

Øvelser

I den følgende opgave er diagrammet med en lineær regression foretaget, og øvelsen går på tolkningen af denne.

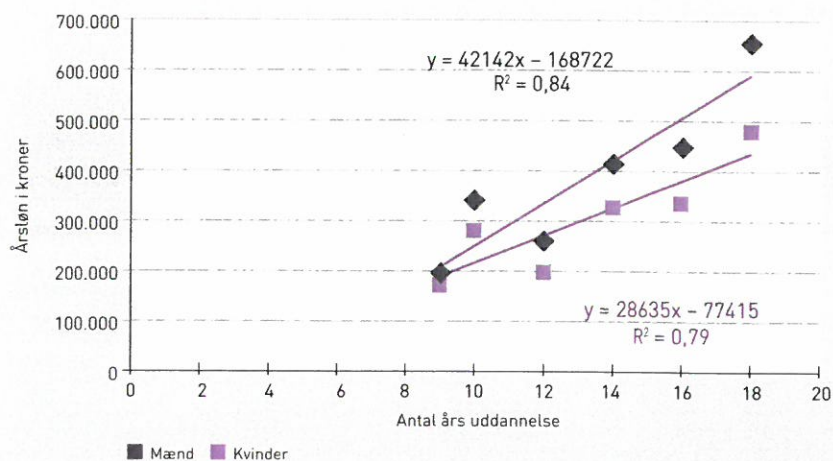
Opgave 3:

Hvad kan der af figur 3.3 udledes om sammenhængen mellem årsløn og antal års uddannelse samt forskelle i løn for henholdsvis mænd og kvinder i Danmark?

Du skal anvende viden om ligestilling og ulighed.

Figur 3.3:

Diagram med en lineær regression, der viser sammenhængen mellem antal års uddannelse og årsløn i kroner for hhv. kvinder og mænd



Kilde: Undervisningsministeriet. Samfundsfag A, 30. maj 2016 (med visse ændringer i opgaveformuleringen).