Multipel regression i matematik

Introduktion

Dette materiale er tænkt som en inspiration til hvordan matematiklærere kan gribe multipel regression an i STX på matematik A-niveau og er særligt relevant for elever med en samfundsfag A, matematik A studieretning. Det giver god mening at afholde modulet før end det samfundsfaglige modul. Først præsenteres sammenhængen mellem lineær og multipel regression og dernæst gives der introduktion til hvordan man i praksis kan lave sin regression i Excel, samt hvordan man kan tolke på de fremkomne modeller.

Fra lineær til Multipel regression

Denne del knytter sig til videoens første del frem til 7:36 og er en kort introduktion til arbejdet med multipel regression i matematikundervisningen.

Forudsætningen for at arbejde med multipel regression er at man har arbejdet med den lineære regression. Langt hen ad vejen er multipel regression en naturlig udvidelse herfra.

I den lineære regression arbejder med funktioner af typen

$$y = ax + b \quad (1)$$

Og forsøger at finde den bedste rette linje gennem datapunkterne (x_i, y_i), hvor indeks *i* angiver, at vi har en række datapunkter.

Med multipel regression arbejder vi i stedet med funktioner af typen

$$y = a_0 + a_1x_1 + a_2x_2 + \cdots + a_nx_n \quad (2)$$

Her er der altså en model med op til n forklarende variable. Disse kaldes til tider også *kovarianter*. Fordel er at vi med denne mere komplekse model forestiller os, at vi kan beskrive mere komplekse problemstillinger, der ikke kun afhænger af en enkelt forklarende variabel.

På samme måde som man med en lineær model snakker om hældningskoefficienter og skæring med y-aksen kalder vi a_0 for skæringen med y-aksen og $a_1 \ldots a_n$for hældningskoefficienter. Ligeledes kan vi beregne kvadratsummen, der forkortes SSE (Sum of Squares Errors) med følgende formel:

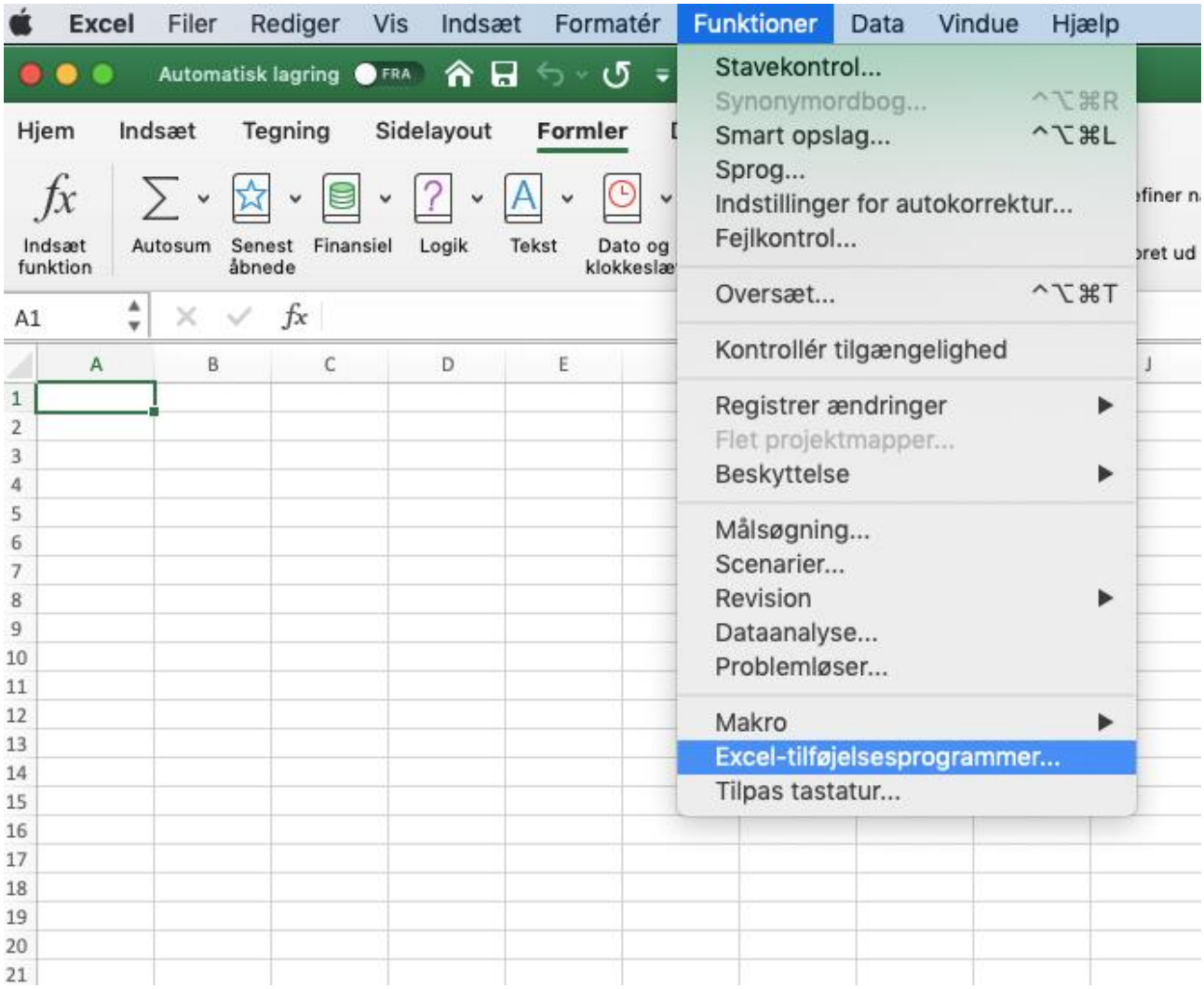
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (3)$$

Hvor y_i angiver datapunkterne og $\hat{y}_i$ angiver den i'te modelværdi beskrevet ved forskriften i ligningen (2). Med regressionen finder vi altså de koefficienter $a_0 \ldots a_n$, der gør at kvadratsummen bliver mindst mulig. Når vi har optimeret kvadratsummen, har vi også fundet den forskrift, der gør at forskellen mellem vores observationer (datapunkter) og vores modelværdier er mindst mulig.

Vejledning til multipel regression i Excel

Denne del knytter sig især til videoen fra 7:36 og frem. Det beskrives, hvordan man kan lave multipel regression med Excel, samt hvad man kan se efter i outputtet fra Excel. Til slut gives der et datasæt med tilhørende opgaver, så elever selv kan forsøge sig med regressionen.

For at lave multipel regression i Excel skal man bruge et til tilføjelsesprogram der til dataanalyse. Dette kan man tilføje under funktioner, Excel-tilføjelsesprogrammer og så vælge 'Analysis ToolPak'.



Har man en anden version af Excel end denne til mac-brugere kan det være man skal finde tilføjelsesprogrammet ad en anden sti. Søg eventuelt på nettet på 'Analysis ToolPak' og tilføj dit styresystem.

Ønskes det at lave regressionen i et cas-værktøj i stedet har Maple også mulighed for at lave multipel regression. Dette kan man gøre ved at benytte kommandoen 'MultiLinReg'-kommandoen fra 'Gympakken'.

I videoen forklares det ud fra datasættet i bilaget, hvordan man i praksis kan lave en multipel regression.

Hvornår er min regressionsmodel tilstrækkelig god?

I statistik er spørgsmål som disse ofte lettere at stille end at besvare, da vi ikke arbejder med absolutte sandheder. Men herunder kan der gives nogle punkter man kan overveje.

Et sted man kan begynde, er at overveje sin undersøgelse som helhed. Er det en relevant undersøgelse? Er der rimelig grund til at antage at de forklarende variable kan have en indflydelse på den afhængige variabel? Er data indsamlet på en forsvarlig måde? Dette arbejde ligger udenfor matematikken, men er ikke desto mindre vigtige.

I eksemplet i videoen ser man desuden, at der ikke er en uafhængighed mellem de forklarende variable. Det kaldes *multicollinaritet* og skal undgås, når man vælger sine bud på forklarende variable.

I vurderingen af selve dataanalysen, kan man begynde med at betragte sine koefficienter og konfidensintervaller. Hvis man ønsker at undersøge, hvordan størrelsen på timeløn afhænger af en række forklarende variable herunder uddannelse, ville det undre en at få en negativ hældningskoefficient, da tolkningen så ville være at uddannelse påvirker timelønnen negativt. Resultatet kan skyldes at modellen ikke er tilstrækkelig god til at beskrive data og behøver ikke nødvendigvis betyde at uddannelse har en negativ indflydelse på en forventet timeløn. Som minimum bør man overveje om der findes andre forklaringer eller om modellen er tilstrækkelig god.

Dette gælder også hvis et konfidensinterval for en af hældningskoefficienterne indeholder tallet nul. Tolkningen af dette tilfælde kan jo således være, at der ikke er tilstrækkelig med belæg til at vurdere, at den forklarende variabel har en betydning for forudsigelsen af den afhængige variabel.

Se desuden på p-værdien, der som udgangspunkt tester med et 5% signifikansniveau. For at være tilfreds med modellen kan man se om "Signifikans F" er mindre end 0,05. I så fald vil der være tegn på at nogle af de forklarende variable synes at kunne bruges til modellen. Der regnes også p-værdier knyttet til de enkelte hældningskoefficienter. Hvis vi ikke skal forkaste modellen, skal disse ligeledes være mindre end 0,05. Får man overordnet set en acceptabel p-værdi, men har en eller flere forklarende variable med en for høj p-værdi kan man forsøge at lave en ny model, hvor man har fjernet de forklarende variable, der havde en for høj p-værdi.

Forklaringsgraden eller determinationskoefficienten betegnes R^2 og giver forholdet mellem den forklarede variation og den totale variation.

$$R^2 = \frac{SSR}{SST} \quad (4)$$

Hvor vi har SSE beskrevet tidligere i ligning (3) kan den forklarede variation SSR beregnes som

$$SSR = \sum_{i=1}^{n} (\hat{y_i} - \bar{y}))^2 \quad (5)$$

Hvor $\hat{y_i}$ angiver den i'te modelværdi og $\bar{y}$ angiver gennemsnittet af samtlige y-værdier fra datasættet. SST kan beregnes som summen af SSR og SSE, det vil sige $SST = SSR + SSE$.
R^2 varierer mellem 0 og 1 og vi ønsker for den gode model med en høj forklaringsgrad at få en forklaringsgrad tæt på 1. Vi bør dog overveje, at jo flere forklarende variable der medtages, jo højere vil forklaringsgraden som regel blive. Derfor kan det give mening at bruge den justerede forklaringsgrad, der tager højde for dette. Formlen for beregning af den justerede R^2-værdi kan ses her.

$$R^2_{justeret} = 1 - \frac{n-1}{n-(k+1)} \cdot (1 - R^2) \qquad (6)$$

Hvor n angiver antallet af datapunkter og k antallet af forklarende variable. Til tider tillægges R^2-værdien eller p-værdierne meget stor betydning, men man skal være påpasselig med at konkludere for firkantet om en model kan godkendes eller ej. Der kan være stor forskel på korrelation og kausalitet og ofte vil man blot kunne konkludere, at man ikke kan forkaste en model eller at den synes at være anvendelig.

Opgaver

Førend man giver sig ud i en stor undersøgelse, kan det være rart at have oplevet, at det kan lykkes at lave regressionsanalysen på et eksempel. Her er der foreslået fire små opgaver, der lægger sig op ad videointroduktionen og som kan gives til elever inden man påbegynder et større tværfagligt projekt med samfundsfag.

Opgave 1

Åben dataarket 'Data timeløn' i Excel og lav scatterplots for de tre forklarende variable. Giv akserne korrekte titler og undersøg om Excel har indtegnet punkterne korrekt ift. hvilken af variablerne der er afhængig og hvilken der er uafhængig.

Opgave 2 –Multicollinearitet

Lav en forskrift for sammenhængen mellem alder som afhængig variabel og uddannelse samt erfaring som forklarende variable. Brug argumentationen fra videoen som inspiration.

Lav herefter en multipel regressionsanalyse og tolk på forskellen mellem regressionen og den forskrift du selv havde gættet på.

Overvej hvorfor de forklarende variable skal være uafhængige af hinanden.

Opgave 3

Lav multipel regression på baggrund af først hele datasættet og derefter med udvælgelse af kun to forklarende variable. Overvej om det er vigtigt at vælge uddannelse og erfaring eller om du ligeså godt kunne have valgt uddannelse og alder som forklarende variable.

Opgave 4

Tjek beregningen af R^2 og $R^2_{justeret}$ved brug af ligningerne (3)-(6). Brug regnearket til at beregne først SSE, SSR. Herefter kan R^2 og $R^2_{justeret}$ beregnes og sammenlignes med resultatet fra opgave 3.

Diskuter om sandhedsværdien af modellen – I hvor høj grad kan modellen bruges til forudsige timelønnen for kvindelige håndværkere i USA i 1981. Kom i den forbindelse ind på R^2værdien samt størrelsen af stikprøven.

Henvisninger

- Markussen, Bo og Rønn-Nielsen, Anders: "Lineær Regression A-niveau", 9. oktober, 2018: https://emu.dk/stx/matematik/om-lineaer-regression-og-statistik-i-gymnasiet (side 44-49)

- Webmatematik, Introduktion til Multipel regression: https://www.webmatematik.dk/lektioner/matematik-a/statistik/multipel-regression/multipel-regression
- Claus Thorn Ekstrøm, Ernst Hansen og Per Bruun Brockhoff, Statistik i gymnasiet, 17. januar 2017: https://emu.dk/sites/default/files/2019-02/Statistik%20i%20gymnasiet%20Ekstr%C3%B8m%20Hansen%20Brockhoff%20%202017.pdf
- Claus Thorn Ekstrøm, Ernst Hansen og Per Bruun Brockhoff, Brugen af R^2 i gymnasiet, 17. januar 2017: https://emu.dk//sites/default/files/2019-02/Brugen%20af%20R2%20i%20gymnasiet%20Ekstr%c3%b8m%20Hansen%20Brockhoff%20%202017.pdf
- Tyler Vigen, Spurious-correlations: http://www.tylervigen.com/spurious-correlations